INSTITUTO CARLOS CHAGAS
MESTRADO EM BIOCIÊNCIAS E BIOTECNOLOGIA

LOUISE ULRICH KURT

AN APPROACH FOR QUALITY CONTROL AND CHARACTERIZATION OF PROTEIN CONFORMERS BY USING CROSS-LINKING MASS SPECTROMETRY AND PATTERN RECOGNITION

CURITIBA
2020

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO CARLOS CHAGAS
Mestrado em Biociências e Biotecnologia

LOUISE ULRICH KURT

An approach for quality control and characterization of protein conformers by using cross-linking mass spectrometry and pattern recognition

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Biociências e Biotecnologia, do Instituto Carlos Chagas, como parte dos requisitos necessários à obtenção do título de Mestre em Ciências em Biociências e Biotecnologia.

Orientador: Dr. Paulo Costa Carvalho
Coorientador: Dr. Fabio Cesar Gozzo

CURITIBA
2020

ABSTRACT

The study of protein structures and conformational changes, according to the cell's physiological state, is fundamental to major biotechnological advances such as the development of pharmaceutical drugs. X-ray diffraction and nuclear magnetic resonance (NMR), the gold standards for structural analysis, present limitations such as: X-ray diffraction requiring the formation of a crystal, which is not always possible, and then requiring the crystal to efficiently diffract; NMR being limited to "small" protein complex; and both methods requiring milligrams of purified protein, and being usually limited to the most abundant conformer in the mixture. To overcome these issues, chemical cross-linking coupled with mass spectrometry (XL-MS) emerges as a prominent method for generating structural data by covalently stabilizing a protein system with cross-linkers. Subsequently, the system is digested, and the covalently linked peptides can be identified by mass spectrometry with SIM-XL, a tool developed by our group. We recently observed the existence of many cross-links not satisfying distance constraints according to crystal counterparts. We postulate these cross-links originate from other conformers in the same sample. Our goal was to provide a method that sheds light on different structures of conformers by referring to systems exposed to different biological conditions, analysing the contents with cross-linking mass spectrometry, and providing a software to enable the quality control and assessment of the dataset generated and interpretation of these experiments. To achieve this, we used the 90kDa heat shock protein (HSP90) as a study model under four biological conditions and relied on applying unsupervised clustering on extracted-ion chromatograms (XIC) of identified cross-linked peptides to enable characterization of the different conformers; we emphasize this approach is completely new.

Keywords: cross-linking; mass spectrometry; conformer; quantification; bioinformatics.

RESUMO

O estudo de estruturas de proteínas e suas mudanças conformacionais, de acordo com o estado fisiológico da célula, é fundamental para avanços biotecnológicos como o desenvolvimento de novos fármacos. As metodologias de difração em raio-X e a de ressonância magnética nuclear (RMN), os padrões ouro para análises estruturais, apresentam várias limitações como: na técnica de raio-X, a necessidade de formação de um cristal que difrate com eficiência, o que nem sempre é possível; RMN sendo limitada a “pequenos” complexos proteicos; e os dois métodos requerendo miligramas de proteína purificada, sendo as técnicas geralmente limitadas ao confôrmero mais abundante na solução. Para contornar esses problemas, a técnica de *cross-linking* químico associado a espectrometria de massas (XL-MS) emergiu como uma metodologia de alto potencial, utilizando a estabilização de sistemas proteicos usando agentes de *cross-linking* (chamados de *cross-linkers*) na geração de dados estruturais. Brevemente, o sistema proteico é digerido, um *cross-linker* é introduzido na mistura e os peptídeos ligados covalentemente podem ser identificados usando dados de espectrometria de massas no *software* SIM-XL, uma ferramenta de busca desenvolvida pelo nosso grupo. Recentemente, nós observamos a existência de muitos *cross-links* identificados com confiança que não satisfazem restrições espaciais de acordo com estruturas cristalográficas homólogas presentes na literatura. Nós levantamos a hipótese de que tais *cross-links* originam de outros confôrmeros presentes na mesma amostra. Neste trabalho desenvolvemos um método para ajudar na elucidação de diferentes estruturas de confôrmeros ao expor sistemas proteicos à diferentes condições biológicas, analisá-los por XL-MS, e fornecer um *software* capaz de fazer avalições de controle de qualidade do experimento e interpretar quantitativamente os dados gerados. Para alcançar esses objetivos, usamos a proteína HSP90 como um modelo de estudo sob quatro condições biológicas, e aplicamos um algoritmo de agrupamento não-supervisionado em resultados de cromatogramas de íon extraído (XIC) dos *cross-links* identificados para possibilitar a caracterização de diferentes confôrmeros.

Palavras-chave: cross-linking; espectrometria de massas; confôrmero; quantificação; bioinformática.

FIGURE LIST

TABLE LIST

ABREVIATION LIST

ADP – Adenosine Diphosphate
AMP – Adenosine Monophosphate
ATP – Adenosine Triphosphate
CTD – C-Terminal Domain
DSS – Disuccinimidyl Suberate
ESI – Electrospray Ionization
GUI – Graphical User Interface
HSP90 – 90kDa Heat Shock Protein
KDE – Kernel Density Estimation
LC – Liquid Chromatography
MALDI – Matrix-Assisted Laser Desorption Ionization
MD – Middle Domain
MS – Mass Spectrometry
MS1 – Survey scan
MS2 – Fragmented Precursor Ion Spectra
MS/MS – Tandem Mass Spectrometry
NMR – Nuclear Magnetic Resonance
NTD – N-Terminal Domain
PDB – Protein Data Bank
PSM – Peptide Spectrum Matching
QC – Quality Control
TIC – Total Ion Current
XIC – Extracted Ion Chromatogram
XL – Cross-link
XL-MS – Cross-linking Mass Spectrometry

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Protein Structures and Conformations

Proteins are biological macromolecules connected with several biological processes, and as such, are a widely researched field of study. Their functions are dictated by their structures, which can be understood as having multiple levels of organization, referred to as primary, secondary, tertiary and quaternary structures [1] (**Figure 1-1**). The primary structure is the protein's sequential arrangement of peptide-bond connected amino acids. These sequences have an important part in protein three dimensional folding, making the knowledge of the amino acids sequence of utmost importance in structure studies [2].

Secondary structures refer to the local conformations of polypeptide chains that will compose the entire protein structure, such as helices, sheets, turns, and loops. These conformations will be determined according to the interactions between side chains of amino acids, especially due to hydrogen bonds. The most common secondary structures are alpha helices and beta sheets, but other local conformations can also be observed [1,3].

The combination of all secondary conformations into a stable structure form the protein tertiary or 3D structure. While secondary structures represent interactions between nearby amino acids, tertiary structures are established by bonds, namely hydrogen bonds and disulfide bridges, between spaced out amino acids [1,3]. These 3D arrangements can interact between one another, creating quaternary structures. The majority of proteins is composed of these interacting domains, most of them being dimers, that is, two interacting structures [3].

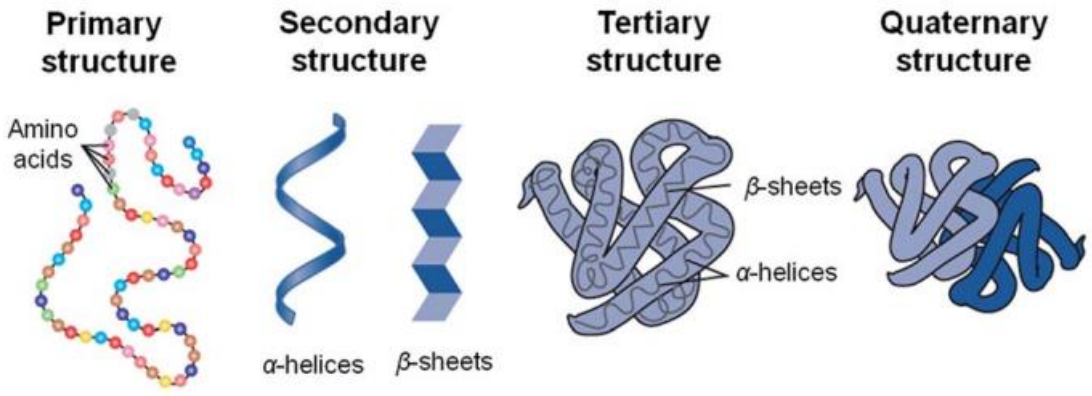


Figure 1-1 – Protein structures. Illustration of the organization levels of proteins, from primary to quaternary. Image Source: Adapted from [1].

The tertiary structure of a protein is of particular interest, as the folding of its chains is directly connected to protein function. A small change in the conformation can render a loss in the biological function [4].

A protein's conformation is also tightly related to its interaction with other proteins. Most protein tasks, such as catalysis, synthesis and degradation of molecules, and transportation are not performed by a single protein, but by a network of inter-protein interactions [5]. Since a slight conformational change can disrupt a whole protein pathway, the understanding of such dynamics is of utmost importance.

Proteins are found in various conformations according to the cell's physiological state. They exist in different structures in a chemical equilibrium, that is, at any given time several conformers may be present in different quantities [6]. The existence of these conformational possibilities within a single sample gives rise to several data analysis challenges as when comparing, say, a control versus diseased state.

While studying protein structures promotes an understanding of their biological activities and is key to solving many challenges, such as drug design, there is still much ground to cover. Public databases statistics provides a view of how much there is to be done; there are around 175 million sequences deposited in UniProt [7], and close to 160 thousand structures deposited in the Protein Data Bank (PDB) [8], that is, fewer than 0.1% of all sequences had their structures determined. Of course, many sequences are still putative or simply not of current biotechnological interest (or are still waiting to have their biotechnological interest discovered), but one of the main reasons for this enormous difference in numbers is the technology available for studying protein structures. The equipment available for protein sequence generation, be it from DNA or from protein samples, can generate more information in less time than it takes to analyse structures [9].

One of the aims of shotgun proteomics is to identify and quantify the proteins found in complex samples such as tissues and biological fluids. Nevertheless, no structural information is provided. To date, the use of methods such as X-ray diffraction and nuclear magnetic resonance (NMR), the gold standards for structural analysis, is not always feasible, for instance, in X-ray experiments, proteins and complexes may produce crystals with poor diffraction or may not crystallize at all. Even then, only the most abundant conformer in the sample is likely to crystallize, leaving precious structural data of other conformational changes in the dark. Moreover, the size of a protein complex is a well-known limitation for NMR experiments. Lastly, both methods

require a considerable amount (milligrams) of purified proteins which is not always achievable [4,10].

To overcome the aforementioned challenges, complementary techniques emerged such as chemical cross-linking coupled with mass spectrometry (XL-MS) [11,12]. XL-MS requires far less sample at lower purity requirements, has a simpler protocol and needs an equipment which does not need to be dedicated exclusively to such experiments [13]. Notwithstanding, XL-ML has been able to provide essential structural information in previous studies, including shedding light on inaccuracies in crystal-derived structures [14], and even full-scale modeling of the structure from a pool of select cross-links. Such methodology is the one used in this project and will be further detailed in the next section.

1.2 Cross-linking Mass Spectrometry (XL-MS)

In our proposed approach, solutions containing the protein/complex under different conditions are covalently stabilized in a reaction with a cross-linker followed by enzymatic digestion. This allows for covalently linked peptides to be identified via tandem mass spectrometry (MS2 or MS/MS) (**Figure 1-2**). The identified cross-linked peptides provide several spatial constraints that reveal important structural information, such as protein folding, topology of complexes, and the interaction region between proteins, among others [12,15].

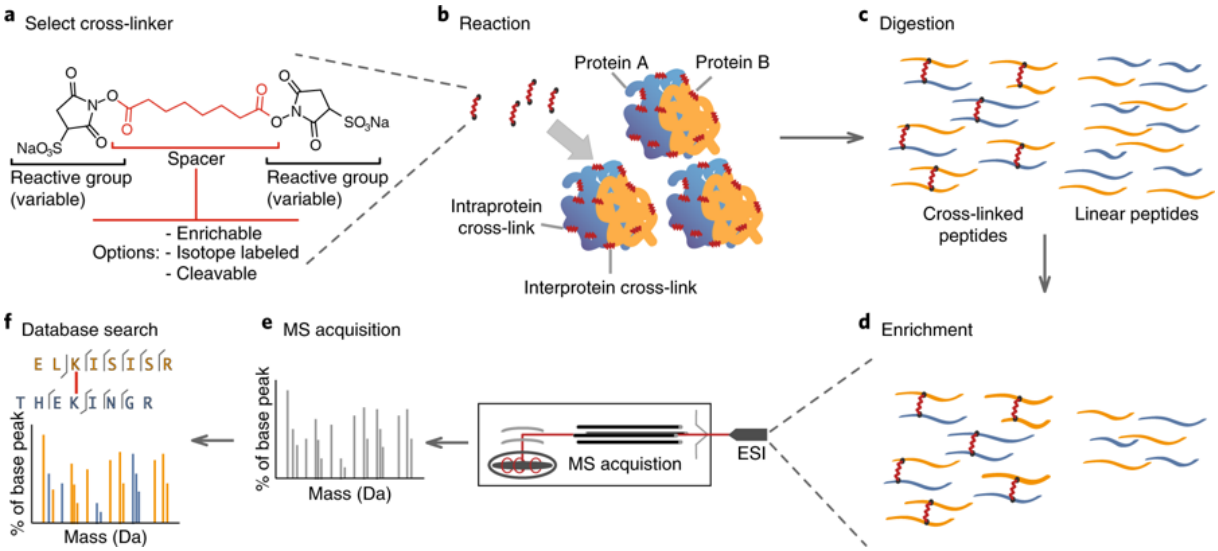


Figure 1-2 – XL-MS. General workflow of a cross-linking mass spectrometry experiment. Image Source: Adapted from [16].

1.2.1 Cross-linking (XL)

Cross-linking consists of creating a bridge-like structure between two molecules, achieved through a third, known molecule, called a cross-linker, that covalently binds to the two molecules. In the case of proteins, the cross-linker will bind to two distinct amino acids that are at a distance within the length of the cross-linker [17].

A cross-linker is composed of two reactive groups in the extremities and a spacer arm in the middle, as shown in **Figure 1-3**. The reactive groups will bind to the specific molecules to be cross-linked, while the spacer arm can consist solely of a hydrocarbon chain to stablish the size of the molecule or they can contain other functional groups with distinctive objectives, such as working as reporter ions or to enrich samples [12,16].

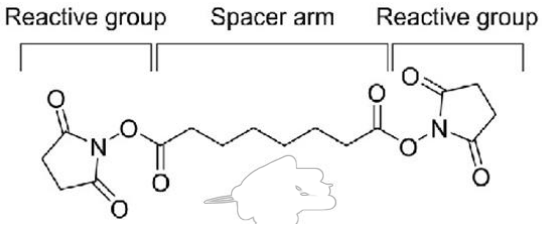


Figure 1-3 – Cross-linker molecule. General structure of a cross-linker, with a reactive group at each end and spacer arm between them. Represented here is the cross-linker DSS. Image Source: Adapted from [18].

Different reactive groups and spacer arms can be used to synthesize a variety of cross-linker molecules, each with its own advantages and specific applications. They are usually classified according to their reactive groups, that is, by the chemical reactivity of the functional groups in the extremities. Some popular cross-linkers for protein conjugation include NH-esters, which react with amine lateral groups in amino acids, carbodiimides, which react to carboxyl groups in amino acids, and photo reactive ones, which activate under UV light and are non-specific, that is, they will bind to any lateral chain in amino acids [12]. Other types of cross-linkers according to their reactive groups can be seen on **Table 1-1**.

Table 1-1 – Common cross-linkers. Most commonly used cross-linkers used for protein studies classified according to their reactive group. Source: Adapted from [19]

Reactivity class	Target functional group	Reactive chemical group
Amine-reactive	-NH2	NHS ester Imidoester Pentafluorophenyl ester Hydroxymethyl phosphine
Carboxyl-to-amine reactive	-COOH	Carbodiimide (e.g., EDC)
Sulfhydryl-reactive	-SH	Maleimide Haloacetyl (Bromo- or Iodo-) Pyridyldisulfide Thiosulfonate Vinylsulfone
Aldehyde-reactive i.e., oxidized sugars (carbonyls)	-CHO	Hydrazide Alkoxyamine
Photo-reactive i.e., nonselective, random insertion	random	Diazirine Aryl azide
Hydroxyl (nonaqueous)-reactive	-OH	Isocyanate
Azide-reactive	-N3	Phosphine

Another way to distinguish cross-linkers is by classifying them as either homofunctional or heterofunctional. Cross-linkers with the same functional group at each end are called homofunctional and will react with amino acids of the same kind, while heterofunctional ones have different reactive groups in their extremities and will react with different amino acids [12].

The spacer arms of cross-linkers will determine the flexibility of the molecule, its solubility and the maximum distance permitted between amino acids that can be bound by it. Longer spacer arms are more flexible, but not recommended for smaller proteins, as the information will not be so specific. Some cross-linkers can be of the zero-length kind, that is, they do not possess a spacer arm and act as an intermediate between the amino acids, not actually being a part of the final bond. Because of that, they will bind only to amino acids that are particularly spatially close to one another (e.g., <~ 5 Å) [20].

The composition of the spacer arm can greatly affect the solubility of the sample. Cross-linkers with spacer arms made of a hydrocarbon chain are very hydrophobic,

and as such require the use of a solvent in the experiment. That makes this kind of cross-linkers more suited for intracellular reactions, as they can more easily permeate the membrane. For *in vitro* essays however, a cross-linker with a hydrophilic spacer arm can be more advantageous, as no solvent will be needed during sample preparation [20,21].

One of the most popular cross-linkers, and the one used in this project, is disuccinimidyl suberate (DSS), shown in **Figure 1-3**. DSS is a homobifunctional cross-linker, with amine reactive groups at each end (NH-ester), reacting mainly with primary amines; for proteins that would include the lateral chain of lysine (K or Lys) and the protein's N-terminal. However it has been shown that NH-esters react to a lesser extent with the side chain of serine (S or Ser) too [22,23]. This cross-linker has a spacer arm composed of a hydrocarbon chain, making it hydrophobic and membrane permeable; it has a length of 11.4Å, so it will only bind amino acids within approximately that distance [24].

After the cross-linkers bind to amino acids, the proteins are digested, forming three different types of links: interlinks, intralinks (or loop-links) or dead-ends (or mono-links) (**Figure 1-4**). Interlinks are formed when a cross-linker binds amino acids from different peptides, that is, they might be sequentially far apart, but are spatially close. In that case, the peptide with the longer sequence will be called α (alpha) and the shorter one will be β (beta). Intralinks bind two different amino acids within the same peptide while dead-ends happen when a cross-linker binds to an amino acid but does not find another reactive site within its reach. While dead-ends do not give much information on structure, they can shed light on regions of the protein that are available for the solvent [16,20,25].

Figure 1-4 – Link types. The three kinds of links that can be formed during a cross-linking experiment. First is a interlink, which happens when two different peptides are linked; second is a intralink, which is when two amino acids of the same peptide are linked and last is a dead-end, which is when a cross-linker binds only to one amino acid. Image Source: Adapted from [20].

The cross-links generated identify spatial restrictions within a protein, so that if amino acid A is bound to amino acid B by a link, then they must be within a distance that is at most the length of the cross-linker. The links formed can be identified using

mass spectrometry, so that information can then be used to infer many things about the studied protein, its structure being one of them.

1.2.2 Liquid Chromatography – Mass Spectrometry (LC-MS)

LC-MS is an analytical method used in many different fields; in biology, typically for the identification of biochemical molecules in the several types of sample, such as proteins, lipids and other metabolic products. The technique works by separating the molecules in the sample, ionizing them, and then measuring their mass to charge ratio (*m/z*) [26]. Analyzing the patterns of these measurements, often in relation to theoretical data is what enables the identification of the chemical components in the sample. In the studies of proteins, LC-MS enables the identification of peptides, thus inferring the proteins present in the sample, as well as the determination of post-translational modifications, interaction partners, 3D structure with the aid of cross-linkers and the quantification of the species present [12,26,27].

Briefly, a typical cross-linking protocol will work as thus: the proteins in solution are stabilized with a cross-linker, then digested by a restriction enzyme such as trypsin, generating countless peptides that will be sent to the LC-MS/MS stage. There the peptides will be separated via liquid chromatography (LC), generally using hydrophobicity as the means for separation. These molecules will then enter the mass spectrometer (MS) continuously. In the MS, the molecules will go from the liquid to gas phase while being ionized in the ion source to then be separated in the analyzer according to their *m/z* ratio that will be read and recorded by a detector [28,29].

Developing new ways to ionize molecules without dissociating them greatly pushed the MS technology forward, particularly with the emergence of techniques such as Matrix-Assisted Laser Desorption Ionization (MALDI) [30] and Electrospray Ionization (ESI) [31]. MALDI can ionize a solid sample into the gas phase and ESI from liquid to gas phase and is, therefore, widely coupled to LC-MS/MS [28]. In ESI, a liquid sample is introduced to the system through a steel capillary in the presence of a strong electric field. This field at the tip of the capillary ultimately vaporizes the solution into highly charged microdroplets, which are, in a typical proteomics case, positive. As the solvent evaporates, the microdroplets size decreases leading to an increased electrostatic repulsion and ultimately to the Coulomb fission. The peptide-ions inside these nanodroplets are eventually “pulled-out” due to the electric field, thus transferring the analyte from the liquid to the gas phase (**Figure 1-5**) [31–33].

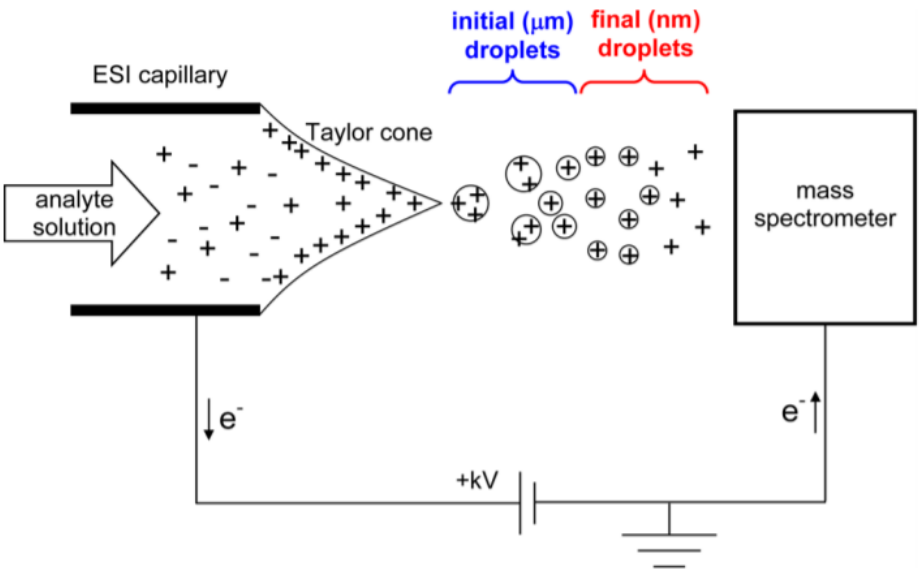


Figure 1-5 – Electrospray Ionization. Schematic diagram of an ESI ion source, with the molecules leaving the capillary and forming nanodroplets which will eventually become ionized molecules in the gas phase before entering the mass spectrometer. Image Source: Adapted from [32]

Once inside the analyser, the molecules will be separated according to their *m/z*. There are many types of analysers, such as quadrupoles, TOF (time-of-flight) and orbitraps. Briefly, quadrupoles work by conducting the ions through an electric field originated from four parallel rods; this field makes the ions oscillate in different ways, with an amplitude directly related to the *m/z* ratio [34]. The limits of this oscillation can be set on the equipment so that only certain ions reach the detector. TOF analysers measure the velocity of ions previously accelerated; the velocity will be inversely proportional to the square root of the ion's *m/z* ratio, so ions with higher *m/z* ratio have the lowest velocity and take more time to reach the detector [28]. Orbitraps focus ions in an elliptical trajectory around an electrode with angular frequencies proportional to their *m/z* ratio, allowing them to be recorded by the detector [34].

The results generated by an LC-MS run are a chromatogram and many mass spectra. Chromatograms overlap the signal from all the ionized peptides that eluted through time, as seen in **Figure 1-6 (A)**. Selecting a single retention time window allows a view of the mass spectral peaks recorded for that given time in a survey scan (MS1) (**Figure 1-6 (B)**). Such peaks, found in the MS1, are not enough to identify a peptide, requiring the use of tandem mass spectrometry (MS/MS). The latter allows the equipment to select a precursor peak of interest from an MS1 scan, isolate the ions with the *m/z* of interest, subject these ions to a dissociation process and obtain a fragment ion mass spectrum (MS2). The process of acquiring an MS1 spectrum, followed by one or more MS2 spectra, is often referred to as a duty cycle. In typical

proteomic experiments, each MS2 spectrum is used to identify a peptide, and in the case of XL-MS, two cross-linked peptides [29,35].

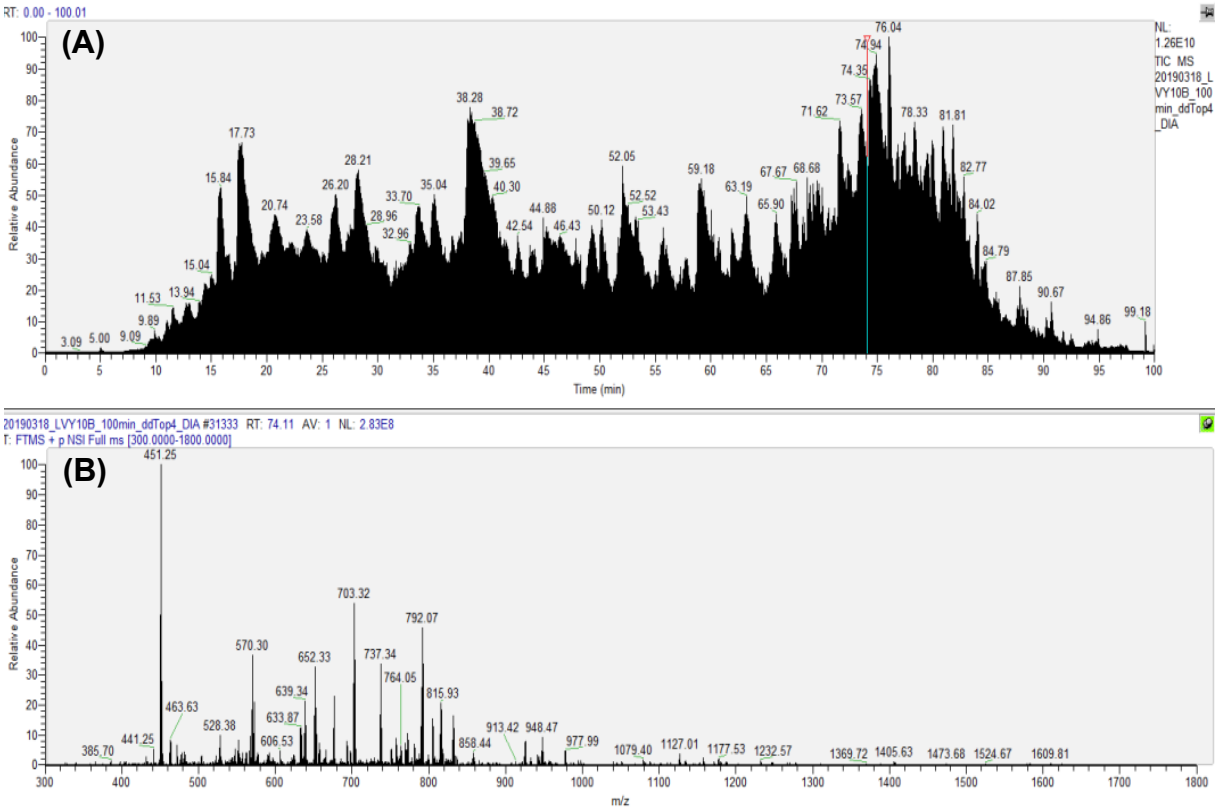


Figure 1-6 – MS experiment resulting data. Representation of a chromatogram (A) and the MS1 scan (B) from the selected retention time as shown in the software Xcalibur (for visualization fo spectra). The intesities of the chromatogram are the summed signal acquired in MS1 spectra for each retention time (what is called Total Ion Current (TIC)).

The identification of cross-linked peptides has roots in an approach known as Peptide Spectrum Matching (PSM). In this approach, the sequences of the proteins being studied are digested *in silico*, that is, a computer software generates all possible peptides and cross-links in the sample and simulates their fragmentation, creating theoretical MS2 scans, which will then be compared with the experimental ones. For XL-MS, the identifications provide spatial constraints that ultimately facilitate the modelling of 3D structures [36,37].

1.2.3 Quantification

There are two widely adopted approaches to quantify peptides by mass spectrometry, the labeled and label free experiments. The former allows multiplexing samples and have them be analyzed in the same MS run; the latter requires each sample to be analyzed independently. In general, for both cases, the quantitative value is relative from one condition to another, and therefore not absolute [38].

This work relies on the label-free approach called extracted-ion chromatogram (XIC), graphically represented in **Figure 1-7**.

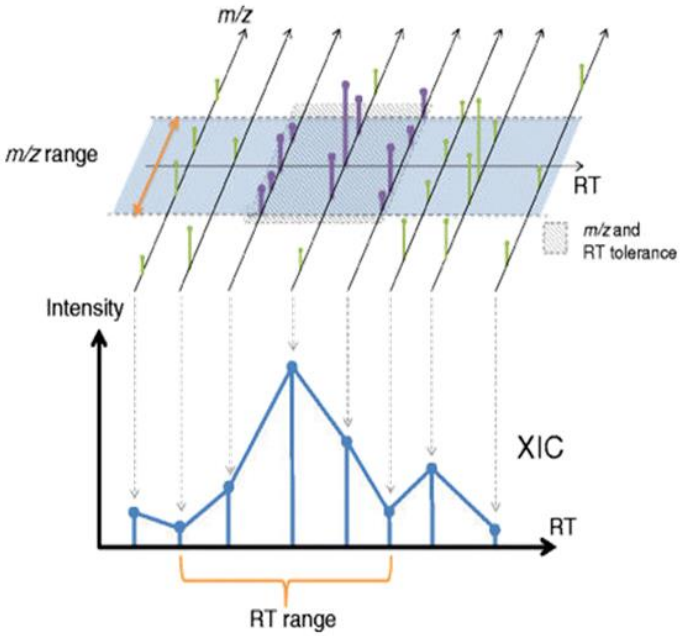


Figure 1-7 – XIC curve. Graphical representation of the extracted-ion chromatogram. The plots on the top are individual MS1 scans, in which a *m/z* range has been selected to be analysed. The bottom plot represents the individual intensities of the peaks selected above plotted against their chromatographic retention time. The XIC will be the area extracted from under this curve. Image Source: Adapted from [39]

In brief, XIC consists of "tracking" a defined *m/z* throughout the chromatography and extracting its elution profile. The MS2 scan allows for peptide identification and thus associating an XIC with a polypeptide sequence. A peptide's precursor *m/z* can be found in various MS1 scans preceding and succeeding its identification (**Figure 1-7);** integrating these intensity-retention points result in quantitative value assigned for that identification [39].

Generating trustworthy quantitative values entails certifying the data with proper quality control (QC) tools. QC also enables assessing the effectiveness of the chromatography process (e.g., reducing the probability of coelution of molecules), in order to improve the quantification process.

1.3 90kDa Heat Shock Protein (HSP90)

We used HSP90 as a study model for this work. HSP90 is one of the most abundant proteins in a cell, comes from a highly conserved protein family, and is one of many molecular chaperones responsible for maintaining proteins correctly folded [40,41]. HSP90 is a homodimer consisting of three domains in each monomer: the N-terminal domain (NTD), where nucleotides bind to the protein; the middle domain (MD), where

some co-chaperones and HSP90's clients interact with the protein; and the C-terminal domain (CTD), involved in the protein's dimerization (**Figure 1-8**) [40,42].

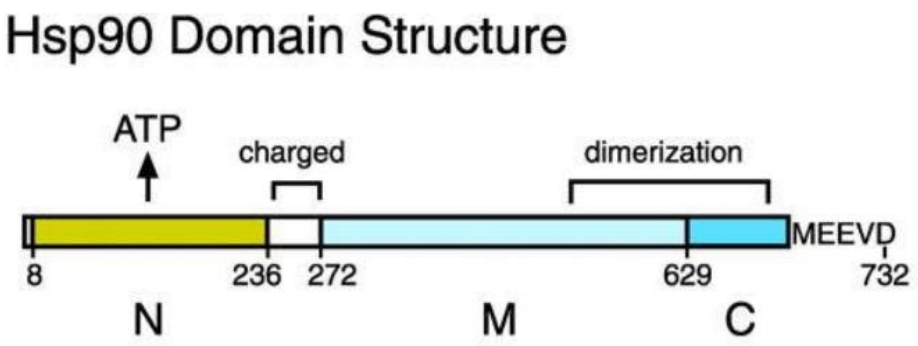


Figure 1-8 – HSP90 domain structure. Representation of the domains in a HSP90 monomer. The N-terminal (N) is where the nucleotides bind and is connected to the middle domain (M) by a charged sequence; the C-terminal (C) is the dimerization site. Image Source: Adapted from [43]

Since the HSP90 acts at the last stage of protein folding, most of its "clients" are already partially folded, which means that HSP90 must alter its conformation to correctly interact with the substrate in question [44] (**Figure 1-9**). This conformational change requires the presence of a nucleotide to activate a dephosphorylation reaction, which gives rise to the different biological states used in this project.

The HSP90 is a flexible protein with a complex conformational cycle, starting with a more open conformation when no nucleotide is bound (called Apo state, dimerized by the C-terminal domain), and shifting to more closed ones according to the nucleotide [45]. When ATP is bound to the protein, it assumes a closed conformation. Being a homodimer, HSP90 presents two pockets for the nucleotides to bind, and it has been shown that while the enzyme still performs its action with only one pocket nucleotide-bound, it has a higher level of activity when both pockets have ligands due to a cooperation between the monomers [46]. This nucleotide-binding process exposes a hydrophobic surface that leads to the dimerization of the N-terminal domain, which in turn causes a rapid hydrolysis of the ATP into ADP, due to the intrinsic ATPase activity of HSP90, slightly changing the protein conformation. When a nucleotide is released, the protein resumes its Apo state [44]. All these conformations have their domains resolved by crystallography; however, this hydrolysis chain can lead to the formation of HSP90 bound to AMP, which does not have a resolved crystal structure.

Structurally speaking, the major changes in the HSP90 conformations come from alterations in the regions connecting the N-terminal to the middle domain and middle to C-terminal domain. The domains themselves remain somewhat conserved, with the

local changes coming mostly from the NTD, which contains a flexible region close to a “lid” formation in the nucleotide-binding pocket, and some changes in the CTD due to the opening and closing of the structures [44,45]. Some experiments have shown that the HSP90 protein appears in an equilibrium between open and closed conformations within the cell and the presence of ATP shifts that equilibrium towards the closed one [44].

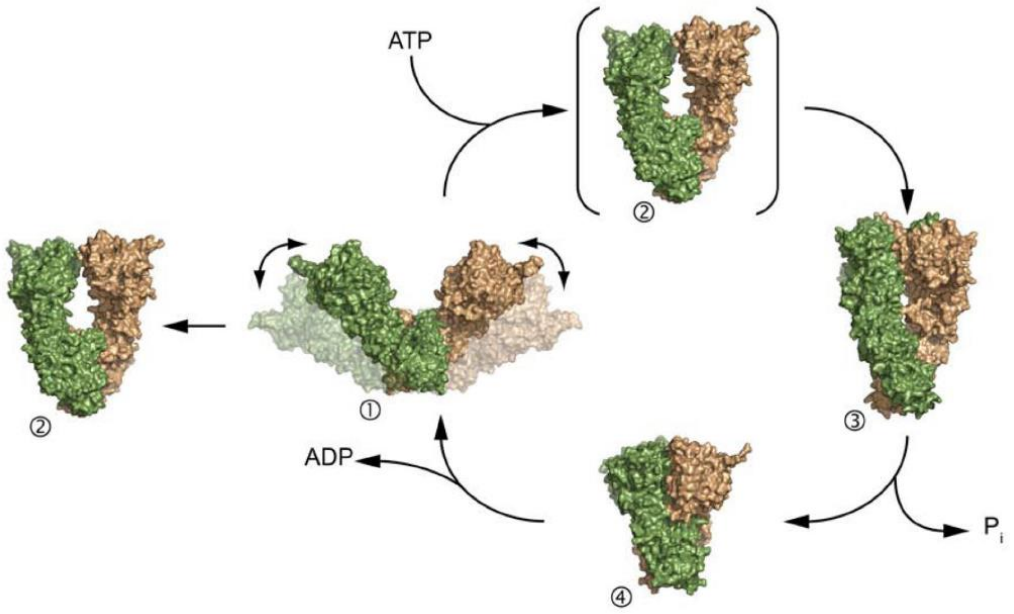


Figure 1-9 - HSP90 conformational cycle. HSP90 conformational cycle, going from nucleotide free (Apo state - 1), to being bound to ATP (2), the going through an intermediate conformation (3), finally hydrolysing ATP into ADP and going to conformation 4. After the ADP is released, the conformation returns to the Apo state. Image Source: Adapted from [44].

Lately, the HSP90 protein has been the subject of many studies, as it is involved in a variety of signalling pathways and can become a target for cancer treatment [45,47,48]. In order to develop new inhibitors for it, the structure and dynamics of the protein must be better understood. Given its importance in the cell and as a drug target, a well-known structure and a complex conformational cycle, the HSP90 makes for an exceptional model for this project.

2. MOTIVATION

The advent of new generation mass spectrometers typically results in the identification of several confident cross-links (distance constraints) that are not compatible with the crystal structure counterpart, which suggests the presence of multiple conformers in solution. This motivated us to pioneer a technology to deal with the quality control and use quantitative cross-linking data to help infer information on protein-complex conformational changes. As far as we know, this is the first computational method tailored towards this goal.

3. OBJECTIVES

Development of an experimental and computational protocol for quality control and quantitative interpretation of data originating from multiple conformers stabilized by cross-linkers and analysed by high-resolution mass spectrometry. The results should shed light on the different conformational configurations of a protein/system.

3.1 Specific Objectives

- Offer an environment in which to enable quality control and assessment of the chromatography and mass spectrometry data obtained;
- Generate a computational platform capable of performing quantitative analysis of XL-MS data;
- Develop a tool that relies on quantitative XL-MS and unsupervised clustering to help give insights on the different conformations of the protein being analysed;
- Use XL-MS data of HSP90 under 4 biological conditions: in the presence of only ATP, only ADP and only AMP, and without any nucleotides bound (Apo form) to validate the methodology;
- Offer graphical interpretation of the results obtained and an environment to easily compare different biological states.

4. COMPUTATIONAL METHODOLOGY AND RESULTS

4.1 Overview

The general workflow of the methodology developed can be seen in **Figure 4-1**. The project was split with focus in two main tools, one for quality control and assessment of mass spectrometry data and one for the analysis of quantitative data for protein conformer elucidation, called RawVegetable and QUIN-XL respectively.

Briefly, an analysis using our pipeline proceeds as follows. First, files generated by the mass spectrometry run go through an assessment on the RawVegetable tool, to ascertain that the experiment ran smoothly, and enough spectra were obtained to enable cross-link identification and extraction of quantitative values. Then the search engine SIM-XL [49] is employed to generate identification files for each biological condition. Such files include the peptide sequences, the number of the MS2 scan used for identification, the XL reaction sites, and a numerical score indicating a confidence level for the identification.

Our software then uses these search results to extract quantitative information from the original mass spectrometry data in the form of XIC. In what follows, the software will then attempt to cluster molecules with similar quantitative profiles. The quantitative profiles identified are then displayed using an user-friendly graphical interface in the form of linear plots.

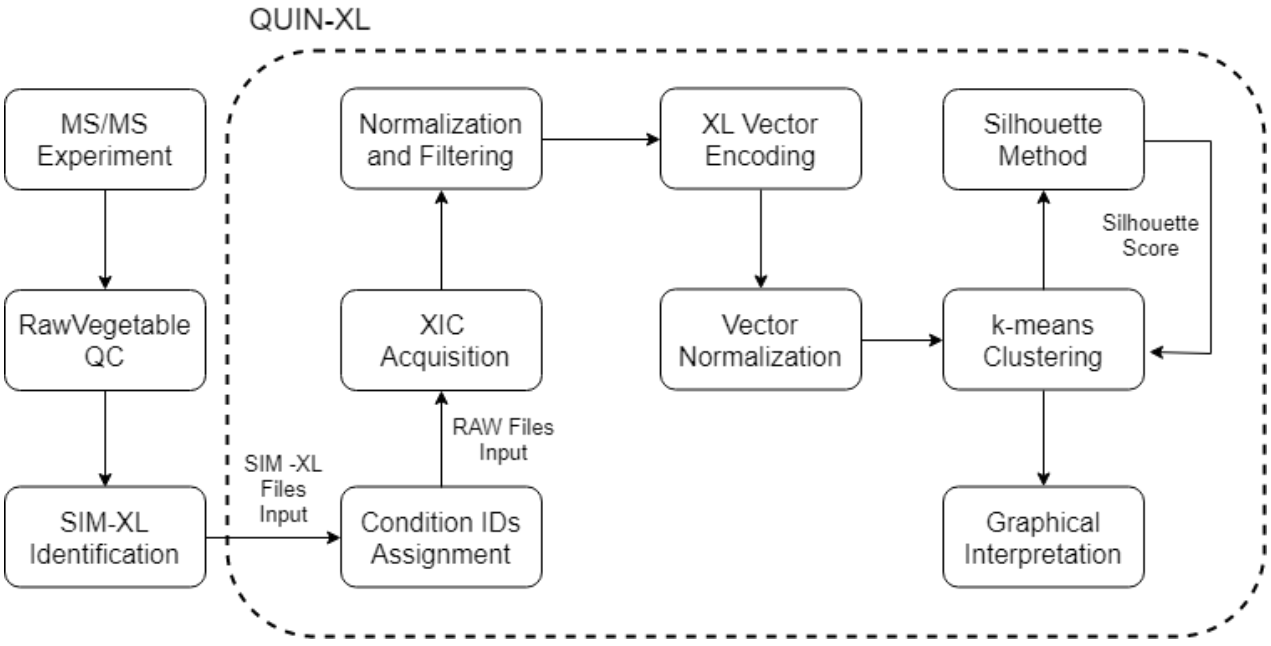


Figure 4-1- Methodology workflow. Workflow of the methodology developed, with the MS experiment files first going through a QC step with RawVegetable, then having the cross-links identified by SIM-XL to finally be quantified and clustered into different possible conformers or regions by QUIN-XL.

4.1.1 Software Implementation

The software was developed using the C# object-oriented programming language, which has numerous available libraries and functionalities, including direct integration with the Windows Foundation Presentation (WPF) subsystem for the development of the graphical user interface. The Integrated Development Environment (IDE) of choice is Microsoft Visual Studio 2019 Community Edition and .NET 4.7.2.

The software shares some of its source code with PatternLab for Proteomics [50], a widely adopted software used for analysing shotgun proteomic data; uses the YADA algorithm [51] for deconvolution of the spectra; the Accord package for *k*-means clustering; OxyPlot for charts; the XCeed WPF package for some of its graphical user interface (GUI) controls; Google's Protobuf protocol for serialization of the project files generated; and the CSMSL library to read .mzML and Agilent files.

4.2 RawVegetable

Chromatography quality control is a critical step in any biological mass spectrometry experiment. Several freely available tools tailored toward shotgun proteomics are available, such as RawMeat (Vast Scientific), which is probably the most widely adopted, but has been discontinued for some years. In analogy to RawMeat, we developed RawVegetable, a tool for general proteomics QC with a focus on XL-MS. RawVegetable includes all RawMeat QC features plus deals with other standard formats such as *.mzML and presents several key unique features presented below.

Description on the use and features of the software is also present on **ANNEX 1**, which is the manuscript of the technical note accepted by the Journal of Proteomics on the 3rd of June 2020.

RawVegetable works by loading mass spectrometry data, such as Thermo RAW files, *.mzML, *.ms1 or Agilent files as input. After the spectra have been loaded, the chromatographic profiles are displayed and all files loaded are listed on the left side, where the user can choose which ones to view (**Figure 4-2**). Once the data have been loaded, there are a few analyses the user can make, such as a charged chromatogram, a TopN distribution, a search for XL artifacts. Loading PatternLab for proteomics [50] *.xic and *.plp or SIM-XL [52] output files enables a reproducibility analysis. All these features are better described in the following sections.

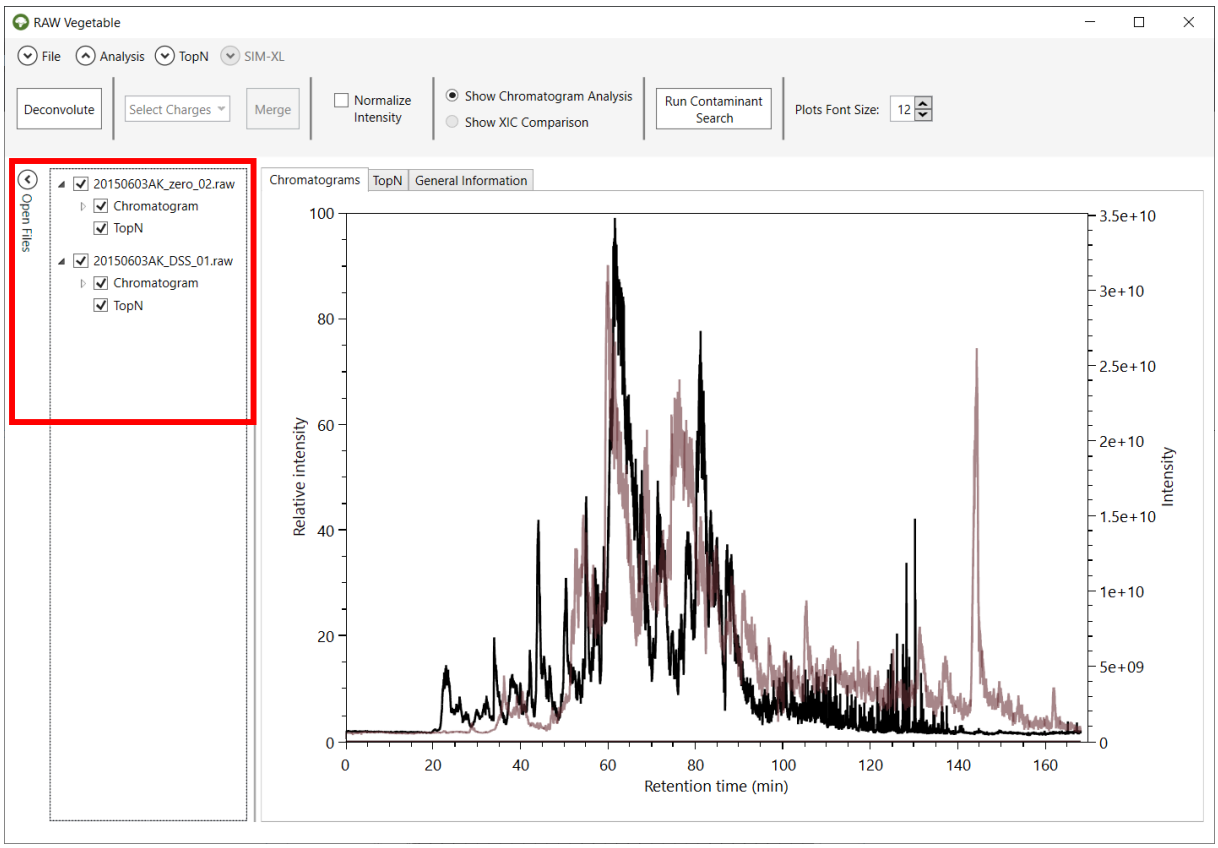
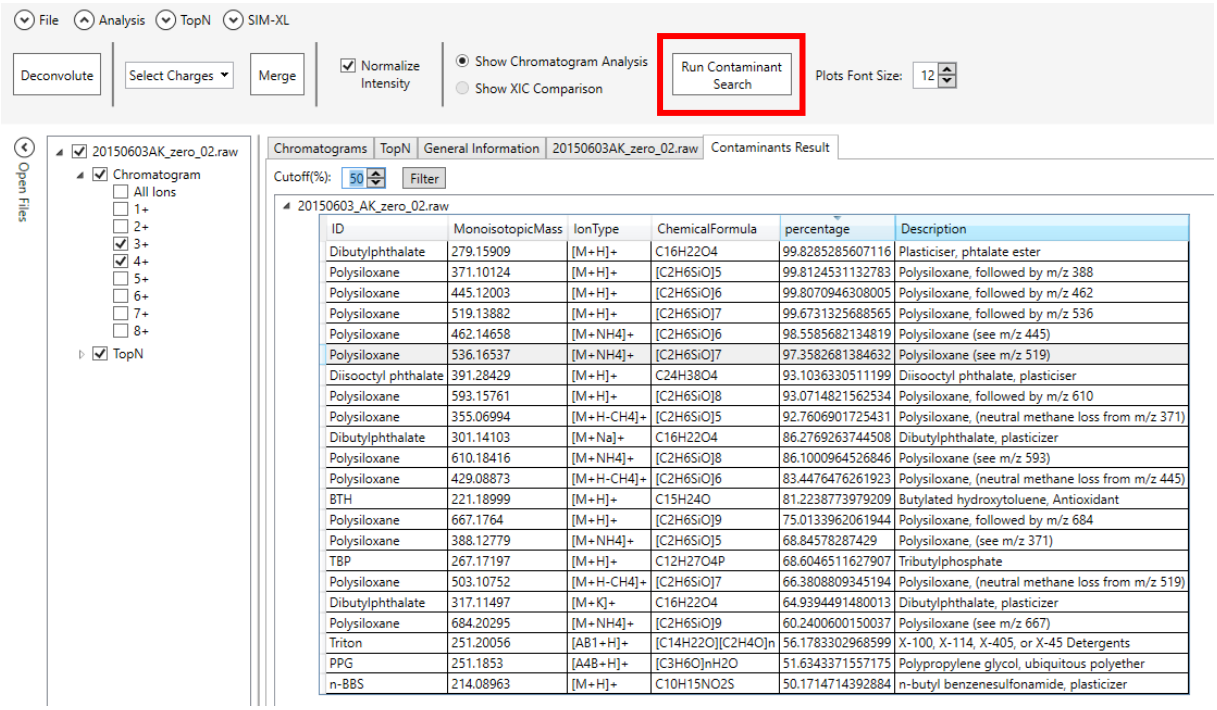


Figure 4-2 – RawVegetable's screen after the spectra file have been loaded. On the left side it is possible to see the menu where the user can check which files they wish to view on the screen.

Another analysis the user can make once the spectra files have been loaded is a preliminary search for contaminants. This is based on a list of contaminants described by Keller et al [53], which lists around 700 common contaminants found in mass spectrometry experiments, both with ESI and MALDI ion sources; this list also has a description of the contaminants and the peaks for which to look in mass spectra. The search RawVegetable performs takes all MS1 spectra in a file and looks for all the peaks listed as contaminants for ESI experiments. This can take a few minutes depending on the number of MS1 spectra in the file and results in a table of contaminants and the percentages of spectra they were found in (**Figure 4-3**). This search can give an idea of possible contaminants and what to look for in a deeper analysis.



ID	MonoisotopicMass	IonType	ChemicalFormula	percentage	Description
Dibutylphthalate	279.15909	[M+H]+	C16H22O4	99.8285285607116	Plasticiser, phtalate ester
Polysiloxane	371.10124	[M+H]+	[C2H6SiO]5	99.8124531132783	Polysiloxane, followed by m/z 388
Polysiloxane	445.12003	[M+H]+	[C2H6SiO]6	99.8070946308005	Polysiloxane, followed by m/z 462
Polysiloxane	519.13882	[M+H]+	[C2H6SiO]7	99.6731325688565	Polysiloxane, followed by m/z 536
Polysiloxane	462.14658	[M+NH4]+	[C2H6SiO]6	98.5585682134819	Polysiloxane (see m/z 445)
Polysiloxane	536.16537	[M+NH4]+	[C2H6SiO]7	97.3582681384632	Polysiloxane (see m/z 519)
Diisooctyl phthalate	391.28429	[M+H]+	C24H38O4	93.1036330511199	Diisooctyl phthalate, plasticiser
Polysiloxane	593.15761	[M+H]+	[C2H6SiO]8	93.0714821562534	Polysiloxane, followed by m/z 610
Polysiloxane	355.06994	[M+H-CH4]+	[C2H6SiO]5	92.7606901725431	Polysiloxane, (neutral methane loss from m/z 371)
Dibutylphthalate	301.14103	[M+Na]+	C16H22O4	86.2769263744508	Dibutylphthalate, plasticizer
Polysiloxane	610.18416	[M+NH4]+	[C2H6SiO]8	86.1000964526846	Polysiloxane (see m/z 593)
Polysiloxane	429.08873	[M+H-CH4]+	[C2H6SiO]6	83.4476476261923	Polysiloxane, (neutral methane loss from m/z 445)
BTH	221.18999	[M+H]+	C15H24O	81.2238773979209	Butylated hydroxytoluene, Antioxidant
Polysiloxane	667.1764	[M+H]+	[C2H6SiO]9	75.0133962061944	Polysiloxane, followed by m/z 684
Polysiloxane	388.12779	[M+NH4]+	[C2H6SiO]5	68.84578287429	Polysiloxane, (see m/z 371)
TBP	267.17197	[M+H]+	C12H27O4P	68.6046511627907	Tributylphosphate
Polysiloxane	503.10752	[M+H-CH4]+	[C2H6SiO]7	66.3808809345194	Polysiloxane, (neutral methane loss from m/z 519)
Dibutylphthalate	317.11497	[M+K]+	C16H22O4	64.9394491480013	Dibutylphthalate, plasticizer
Polysiloxane	684.20295	[M+NH4]+	[C2H6SiO]9	60.2400600150037	Polysiloxane (see m/z 667)
Triton	251.20056	[AB1+H]+	[C14H22O][C2H4O]n	56.1783302968599	X-100, X-114, X-405, or X-45 Detergents
PPG	251.1853	[A4B+H]+	[C3H6O]nH2O	51.6343371557175	Polypropylene glycol, ubiquitous polyether
n-BBS	214.08963	[M+H]+	C10H15NO2S	50.1714714392884	n-butyl benzenesulfonamide, plasticizer

Figure 4-3 – Contaminant search screen. Resulting table from the preliminary contaminant search showing possible molecules that had their respective peaks flagged in more than 50% of the MS1 scans.

Tables with general statistics such as number of MS and MS/MS spectra, duty cycle average time and chromatography total time and histograms showing the distribution of precursors' *m/z* are also available.

4.2.1 Charged Chromatogram

There is a correlation between peptide charge state following electrospray ionization and its number of amino acids; thus, cross-linked peptides will mostly present higher charge states than the linear peptides found in traditional proteomic experiments [7]. Moreover, cross-linked peptides are typically less abundant than linear ones, making optimization imperative. As such, chromatographic optimizations with existing tools are severely limited in the face of XL-MS experiments as their viewers cannot independently account for highly charged peptide ions.

This charge state chromatogram module is tailored towards optimizing chromatography for highly charged ion species by independently plotting the chromatographic profile for each charge state.

In order to do this, all MS1 spectra from each file loaded are deconvoluted using the YADA algorithm [51]. Deconvolution algorithms consist of simplifying a spectrum by merging all isotopic and charge envelopes of a molecule into a single peak.

Isotopic envelopes occur because organic molecules are composed mainly by carbon, which is present in nature with a mass of 13Da instead of 12Da 1% of times.

Since cross-linked peptides can contain many carbons, the chance of some of them being ^{13}C is quite high. This ends up creating various peaks of the same molecule with the mass difference of a neutron between them, as shown in **Figure 4-4**. The fact that this mass difference is known is what allows isotopic envelopes to be used for the determination of the charge of the species represented there, using the following equation:

$$Charge = \frac{Neutron\ mass}{Peak\ 2 - Peak\ 1} \qquad (1)$$

Here, Peak 1 and Peak 2 are the *m/z* values of two consecutive peaks in an isotopic envelope (**Figure 4-4**) and the neutron mass is around 1Da.

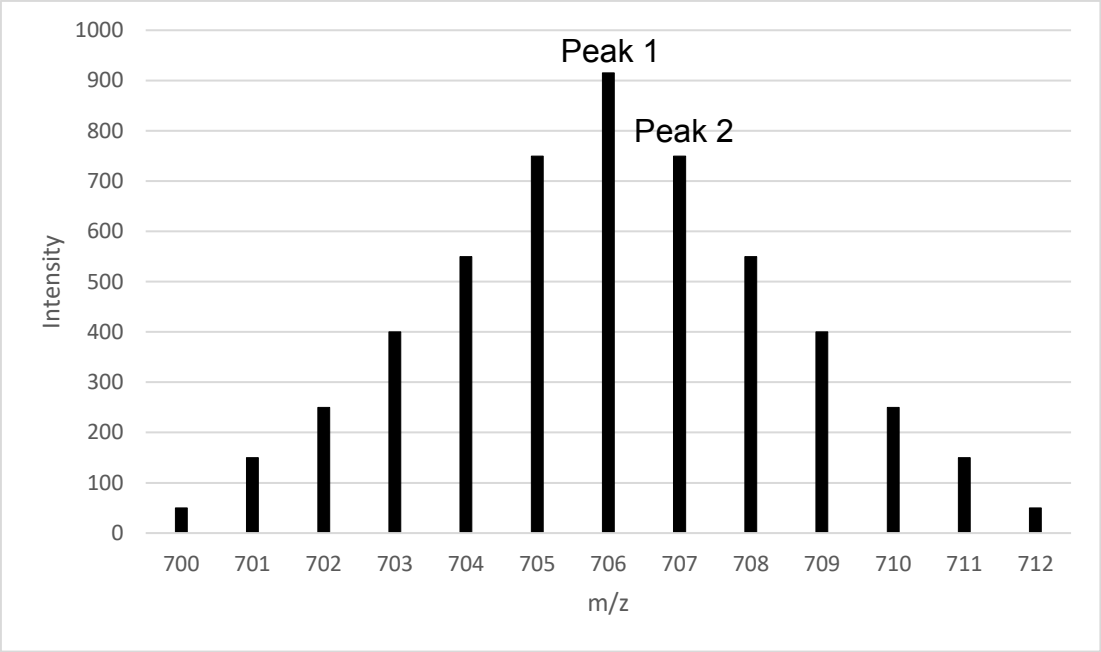


Figure 4-4 – Isotopic envelope. Representation of a isotopic envelope of a molecule with a lot of carbon atoms. Using equation (1), we discover that the charge of this molecule is 1+, so its mass ranges from 700Da to 712Da, depending on the number of ^{13}C it has. In this case, most molecules have six ^{13}C, as the most intense peak is the with is the one that has a difference of around 6Da from the first peak.

This algorithm results in a list of all envelopes found in each spectrum, which contains their respective charge and the summed intensities of all peaks in the envelope. With this information, it is possible to choose a single charge and build a chromatogram using the intensities of all envelopes with the selected charge, as shown in **Figure 4-5**. This new chromatogram can give insights into when most cross-linked peptides are eluting, as they usually appear with charges 3+ and 4+. The chromatography gradient can thus be altered to prioritize such regions, enrich populations of desired charge states, which could help in the acquisition of more MS2 spectra from cross-linked species in lieu of conventional, linear peptides.

In the GUI, the user can then choose which charge chromatogram to view on the left side of the screen, by checking the boxes of each file, as in **Figure 4-5**. It is also possible to normalize the chromatograms on the screen in order to see them all as relative intensities and merge charged chromatograms to see them as one, as in **Figure 4-5**.

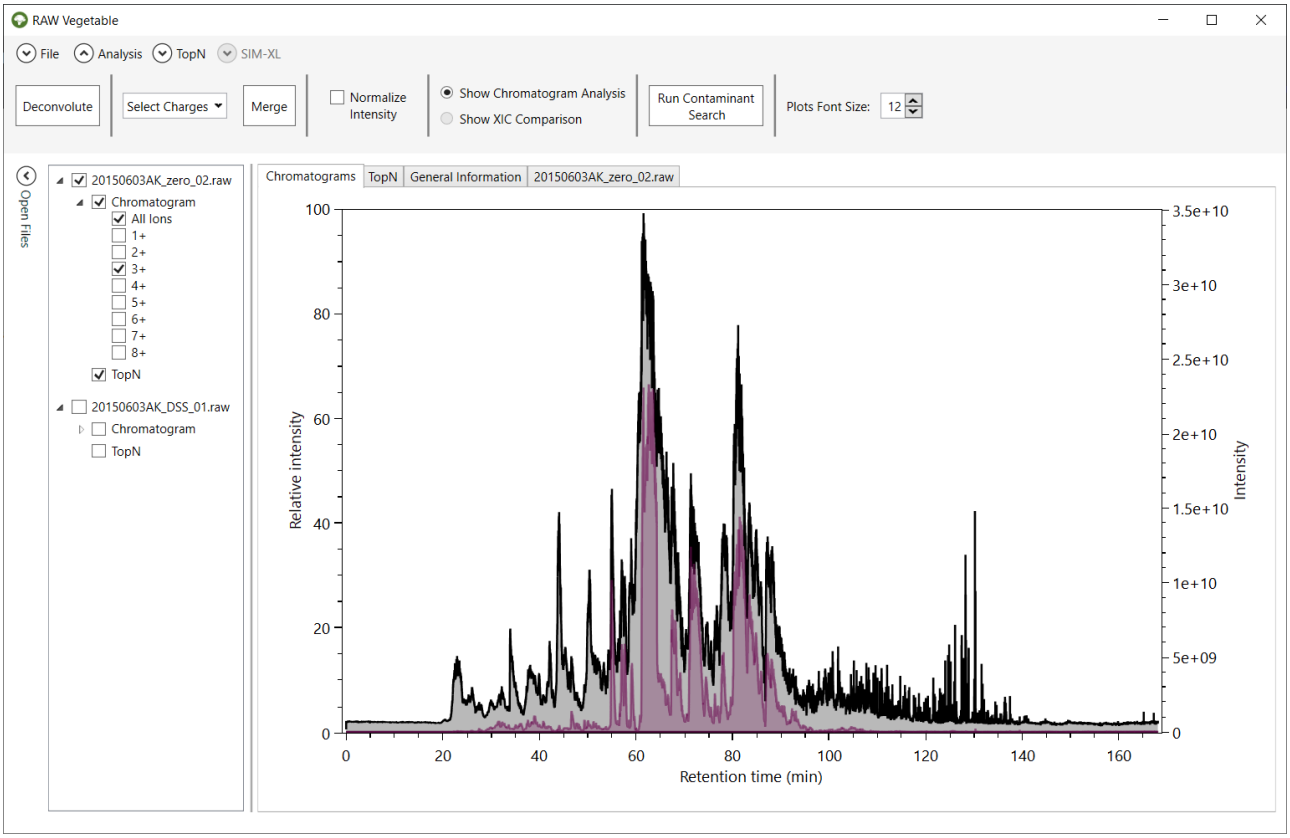


Figure 4-5 - RawVegetable's screen after the deconvolution algorithm has run. Represented in black is the full chromatogram of the file selected in the menu on the left; in purple is the chromatogram considering only species with charge 3+. On the top menu it is possible to select charges to merge, as well as normalize the intensities of the chromatograms.

If the search for cross-links has already been performed, it is possible to identify the regions where most of them appeared by loading the SIM-XL output file into RawVegetable. This will result in all identifications being represented in the chromatograms as small circles at the retention time where the scan was extracted. The user can filter the identification by link type and by the scores calculated by SIM-XL (**Figure 4-6**).

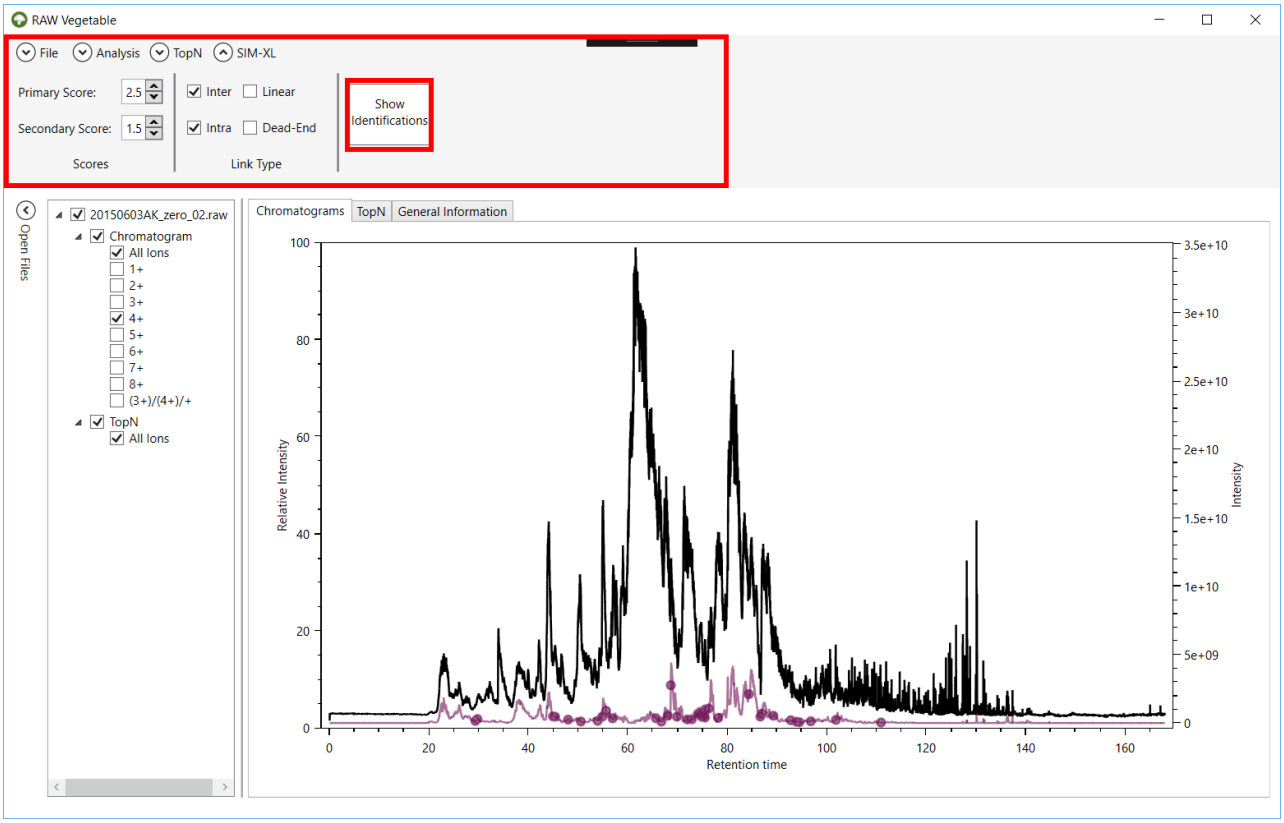


Figure 4-6 – RawVegetable's screen after a SIM-XL output file has been loaded. Here the full chromatogram is represented in black and the charge 4+ is in purple. The purple dots indicate when a cross-link identified with charge 4+ eluted. It is possible to change the score and link type filtering in the top menu.

4.2.2 TopN Distribution

This module shows the density estimation of MS2 scans per duty cycle along the chromatography. This allows pin-pointing retention time intervals that could lead to over- or under-sampling, so the necessary gradient adjustments can be done.

In order to do this, the algorithm applies Kernel Density Estimation (KDE) using a gaussian distribution as the kernel function. Essentially what the algorithm will do is assign a gaussian for each duty cycle, which will then be multiplied by the number of MS2 scans in that cycle. This will result in several gaussians overlapping, which will then be summed, generating a density estimation plot, as in **Figure 4-7**. This way, every gaussian influences all nearby points in the resulting plot, even if minimally.

Given a dataset composed of the retention time of all MS1 scans, that is, the initial retention time of all duty cycles, as the values of x in a plot, the KDE function used to find their respective y value to generate the density estimation plot is the following [54,55]:

$$y = f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (2)$$

Where *n* is the number of duty cycles (which is the same as the number of MS1 scans); *h* is a smoothing factor called bandwidth; $K\left(\frac{x-x_i}{h}\right)$ is the kernel function, where *x* is the retention time of the current point in the KDE being calculated and x_i is the *x* values of the rest of the dataset, that is, the retention time of other duty cycles. The kernel function with x-x_i as parameter represents the influence of all gaussians in the determination of that specific point being calculated, as it returns the intensities of overlapping regions (**Figure 4-7**) to be summed, as represented by the summation (∑) which goes from the first retention time (*i* = *1*) to the last (*i* = *n*).

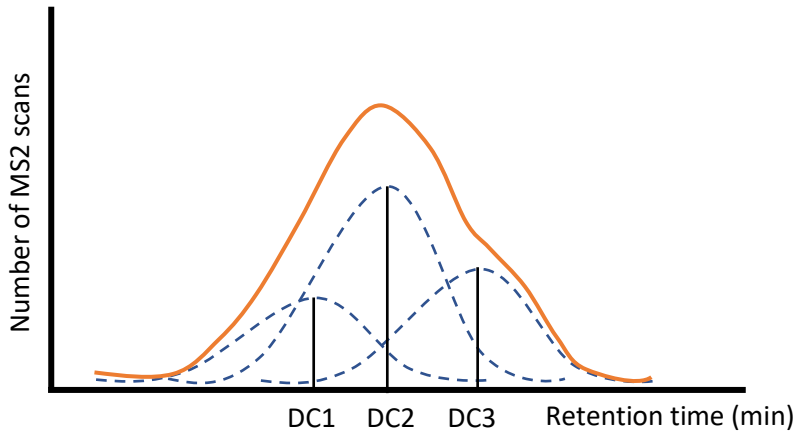


Figure 4-7 – Kernel Density Estimation. Graphical representation of how the Kernel Density Estimation works. For each duty cycle (DC1, DC2 and DC3) a standard normal distribution is built and then multiplied by its respective number of MS2 scans, generating all these different gaussian with overlapping regions, which will then be summed to create a density estimation plot, such as the represented in orange.

The bandwidth (*h*) is a smoothing factor which greatly influences the resulting plot and should be as small as possible without causing the plot to be oversmoothed. It can be a tricky parameter to calculate, but for a gaussian kernel the following equation can be used [56]:

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \qquad (3)$$

Where σ is the standard deviation of the distribution, which is user-defined, but has a default value of 1; *n* is the total number of duty cycles.

The kernel function (*K(x)*) used is a normal (gaussian) distribution with a mean value of 0, which is defined by the following general form [57]:

$$K(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \qquad (4)$$

Where *x* is $\left(\frac{x-x_i}{h}\right)$ as defined by the KDE equation and σ^2 is the variance, which is the square of the standard deviation, and has the same value as in the *h* formula.

The KDE function can then be represented by the following equation:

$$f(x) = y = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left(\frac{x-x_i}{h}\right)^2}{2\sigma^2}} \qquad (5)$$

Applying this function for all duty cycle retention times results in a plot such as the one shown in **Figure 4-8**. This plot permits optimizing the chromatography gradient so that the TopN (MS2 scans in a duty cycle) distribution is as wide and homogeneous as possible through time. This optimization can also help shortening the extremities of the chromatography run, thus shortening the time of the experiment.

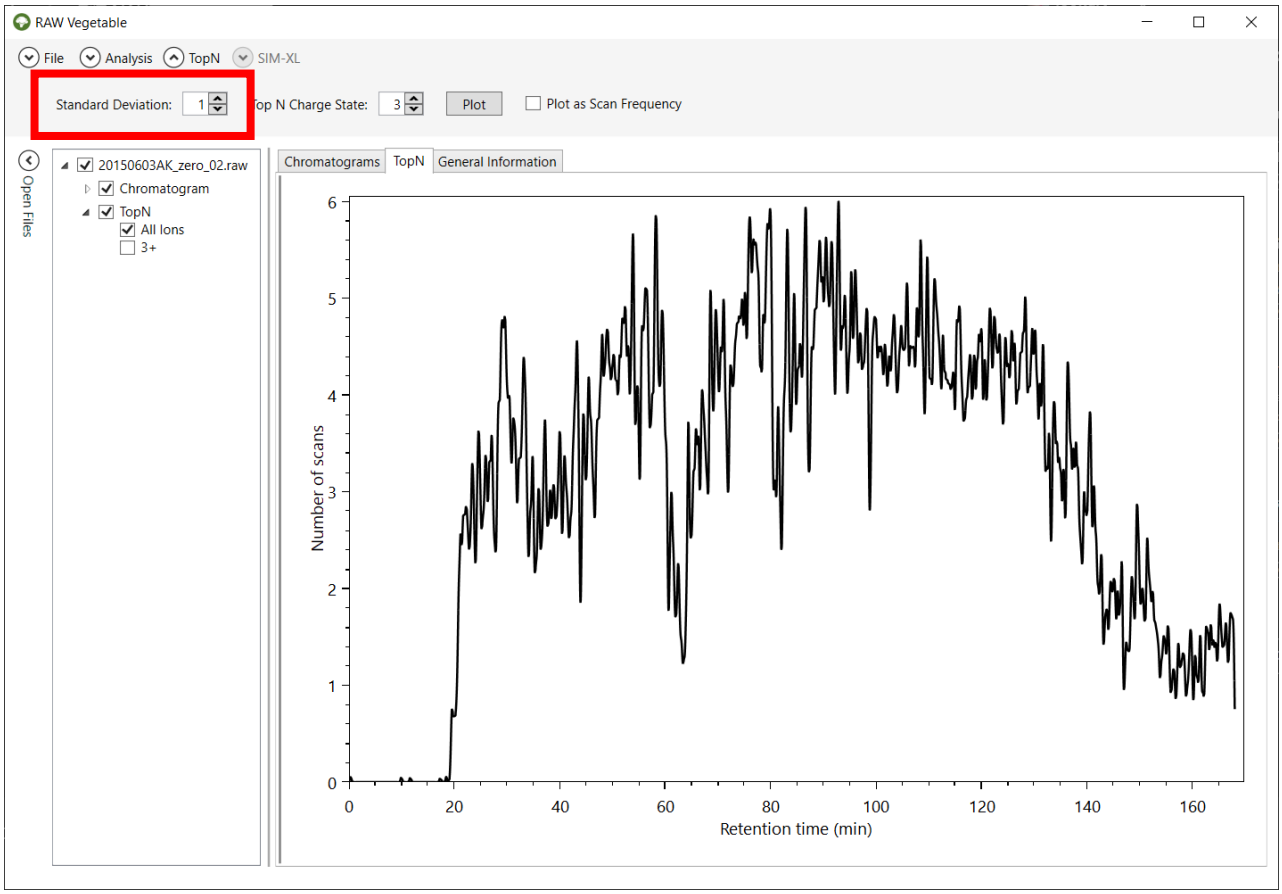


Figure 4-8 – TopN Distribution screen. RawVegetable's screen after the TopN distribution has been calculated, with the user being able to choose the standard deviation to be used during the calculation. In this case the maximum number of scans in a duty cycle was six spectra and that the equipment did not maintain efficiency throughout the run.

It is also possible to calculate a KDE using only MS2 scans with precursors of a specific charge (**Figure 4-9**), to have a better notion of where heavily charged species elute more frequently, similar to the charged chromatograms.

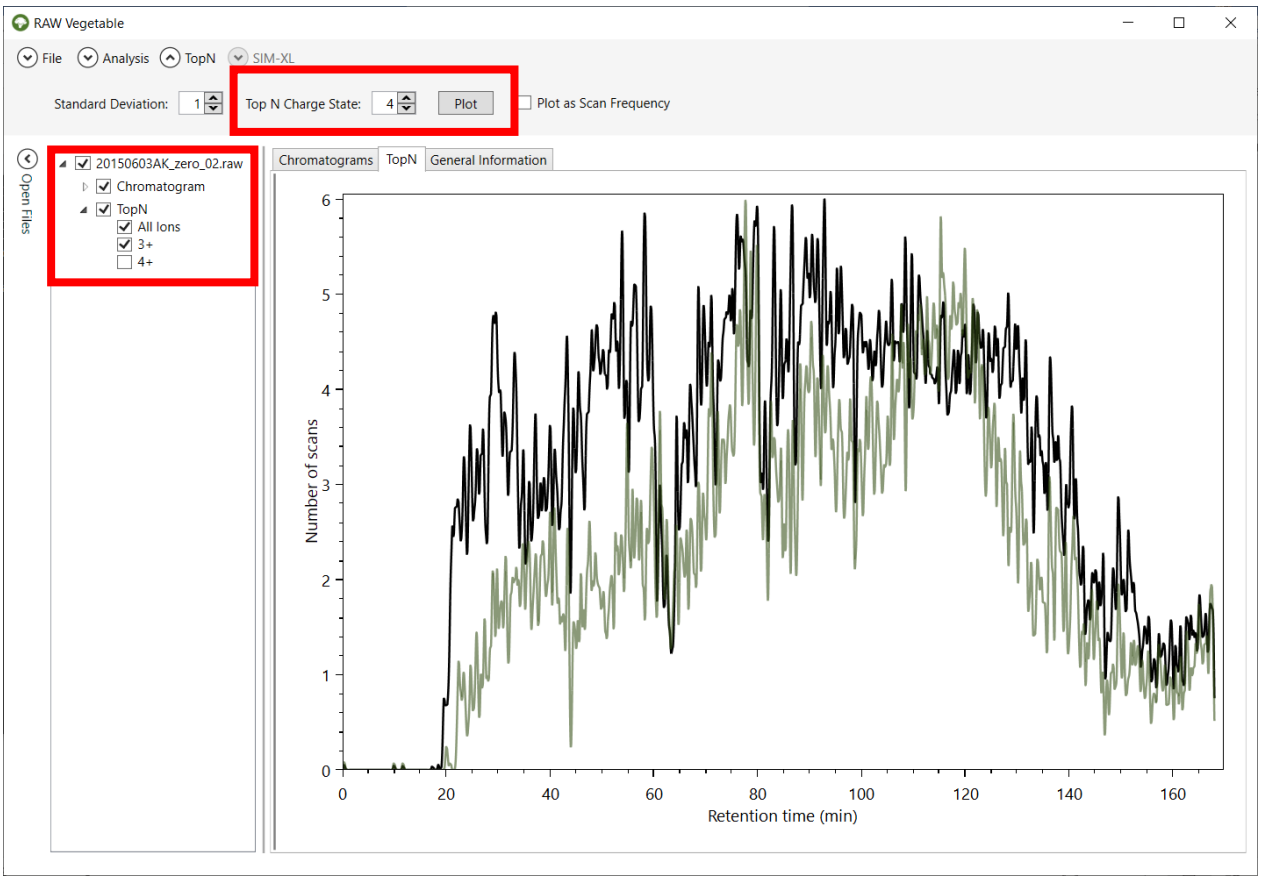


Figure 4-9 – TopN distribution with specific charge state selection. It is possible to view the TopN distribution of species with a single selected charge, as shown by the green curve, which represents the distribution for charge 3+ precursors; in black is the total TopN distribution.

For experiments where the TopN number has not been set and instead only the time for the duty cycle has been specified in the equipment, it is interesting to look at the performance of the run by viewing the TopN distribution as a frequency plot (**Figure 4-10 (A)**). Zooming in regions of interest shows exactly how many scans were obtained in specific duty cycles (**Figure 4-10 (B)**). It is also possible to compare the density estimation with the frequency plot by checking both on the menu on the left side of the screen.

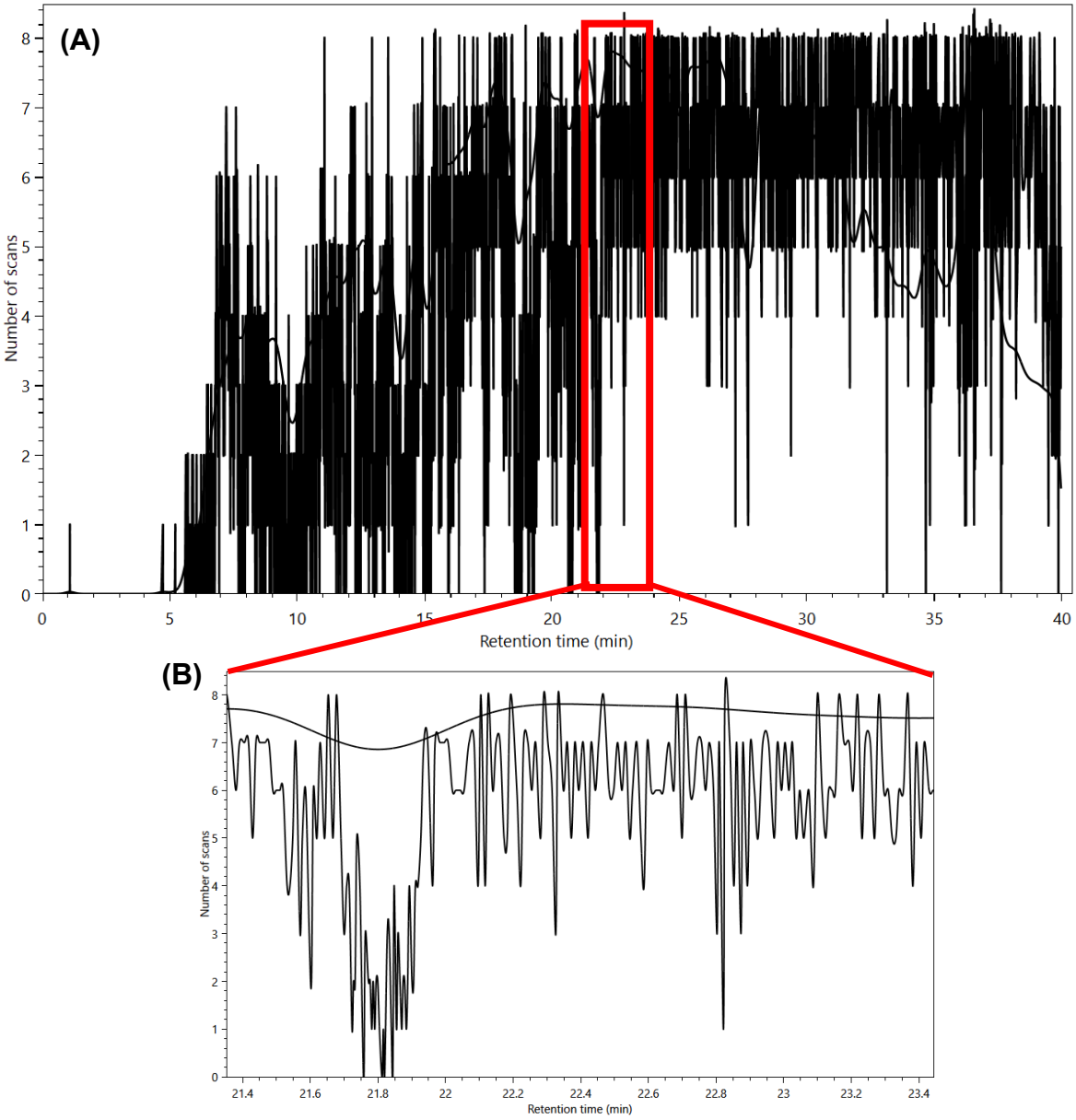


Figure 4-10 – TopN frequency plot. (A) Frequency plot of the MS2 scans during the experiment. (B) Zooming in shows that this plot computes absolute values for the number of scans instead of a distribution as was the case before.

4.2.3 XL-Artifact

Recently, Giese et al. [58] described the presence of some high scoring identifications wrongly taken for interlink peptides owing to the non-covalent dimerization of a linear peptide and an intralinked one during the ionization step of the mass spectrometry, when the molecules are going to the gas phase. Based on this, we added a feature which flags possible species such as these, which we called XL-Artifacts.

RawVegetable uses a SIM-XL output file to identify whether an artifact is at hand by verifying if XICs of all possible separate species (linear α-peptide and intralinked β-

peptide and the inverse) are found during the same retention time range with charges 2+ and 3+. A score is then assigned to the possibly wrong inter-linked XL. This score is calculated as follows: if a linear α-peptide and an intralinked β-peptide or an intralinked α-peptide and a linear β-peptide of any charges have their XICs extracted successfully, one point plus a bonus based on the relative intensities of the XICs are summed to the score. If all four species have valid XICs, then an extra bonus is given. Based on this scoring system, cross-links with scores higher than 1 should already be looked at more closely, however tests made with proteins of known crystallographic structure showed that species with scores higher than 2.5 are the ones more prone to be artifacts.

The result of this algorithm is a table with all interlinked peptides found, with their respective XL-Artifact score and SIM-XL scores listed. The user can also view these possible artifacts represented in the charged chromatograms as red dots, as shown in **Figure 4-11**.

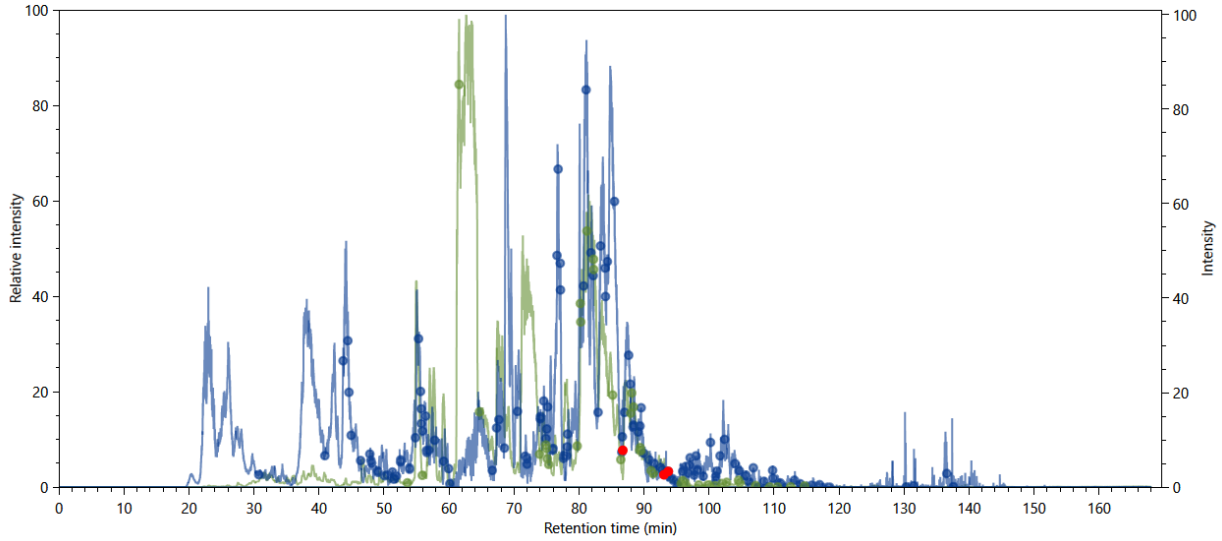


Figure 4-11 – Charged chromatograms with artefact. A charge 3+ (in green) and charge 4+ (in blue) chromatograms. The green and blue dots indicate where cross-linked species were identified with charges 3+ and 4+, respectively; the red dots indicate possible XL-artifacts that should be closely analysed.

4.2.4 Reproducibility Analysis

This module performs all pairwise comparisons of XICs between identifications of different runs, which allows a general view of the chromatographic reproducibility for all runs and therefore to quickly locate problematic MS run.

In order to perform this analysis, RawVegetable needs the PatternLab's *.xic or *.plp result, or the SIM-XL output files as input. The software will then group all common

peptides between the files and proceed to compare their XICs (which should already have been calculated and stored in the input files). This comparison between common peptides is made for each pairwise combination of files, with the XICs of one file representing the *x* values and the XICs of the other file, the *y* values. With these values in hand, it is possible to calculate a similarity score between the files; the user can choose the following metrics to use for the score: r^2, *k* or k^2.

The r^2 metric is the coefficient of determination and calculates how well a set of points fits a linear regression; in this case, a line represented by the equation $y = x$. The general formula for r^2 is the following [59]:

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (6)$$

Where y_i the actual value of *y* being used; $\hat{y}_i$ is the predicted value of y_i, following the equation $y = x$ and $\bar{y}$ is the mean of the *y* values. In this case, the closer the r^2 is to 1, the more similar are files being compared.

The *k* metric is the Euclidean distance between the XIC values of each file. The XIC values of the first file will be represented by the vector *x*, where $x = (x_1, x_2, \ldots, x_n)$ and the second file by *y*, where $y = (y_1, y_2, \ldots, y_n)$. First the values are normalized to be within a range of 0 to 1 and then the Euclidean distance can be calculated using the following equation [60]:

$$k = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (7)$$

Using this metric gives an information which is the inverse of r^2; the closer the *k* value is to 0, the more similar are the files.

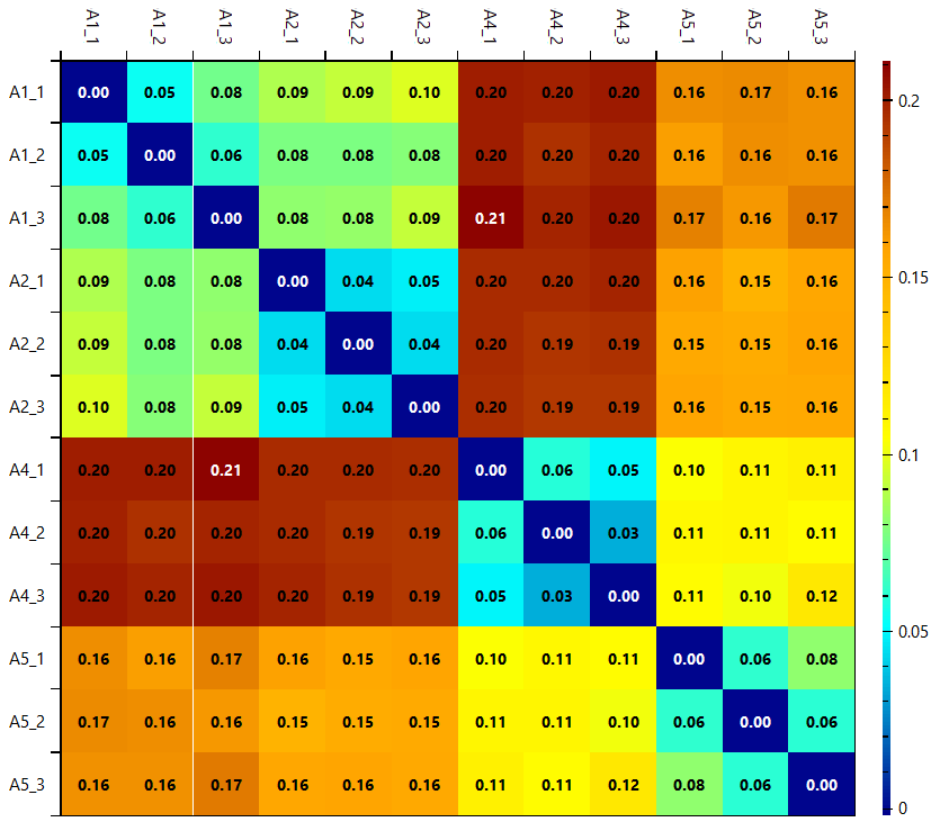
Another metric that can be used is the square of the *k* value (k^2), which is calculated the same way, but accentuates the similarities between files.

The calculation of these scores produces a heatmap comparing files using the chosen metric. The plots generated using the different score systems are shown in **Figure 4-12**.

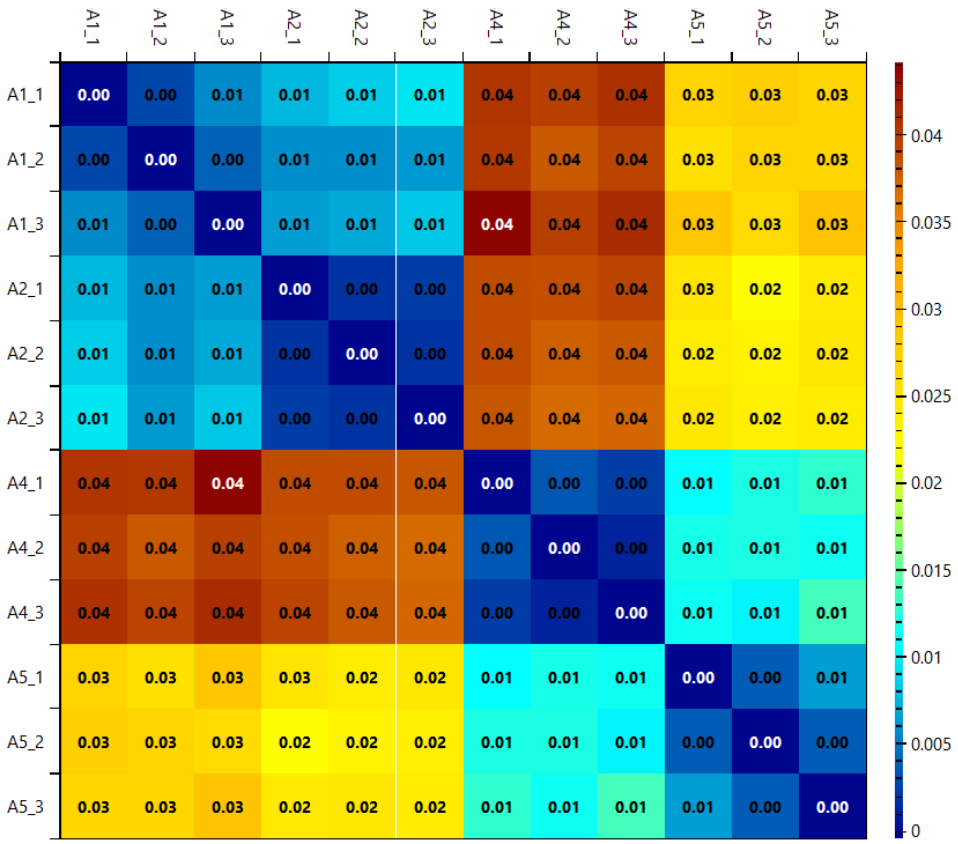
It is also possible compare two specific files by marking them on the left side of the screen. This will result in a dot plot of all common peptides between the files according to their XIC values, along with a trend line and the score value, as in **Figure 4-13**.

This analysis allows to easily pinpoint problematic experiments and determine whether a replicate should be used during the research or not.

(A)


A1_1 A1_2 A1_3 A2_1 A2_2 A2_3 A4_1 A4_2 A4_3 A5_1 A5_2 A5_3
A1_1 0.00 0.05 0.08 0.09 0.09 0.10 0.20 0.20 0.20 0.16 0.17 0.16
A1_2 0.05 0.00 0.06 0.08 0.08 0.08 0.20 0.20 0.20 0.16 0.16 0.16
A1_3 0.08 0.06 0.00 0.08 0.08 0.09 0.21 0.20 0.20 0.17 0.16 0.17
A2_1 0.09 0.08 0.08 0.00 0.04 0.05 0.20 0.20 0.20 0.16 0.15 0.16
A2_2 0.09 0.08 0.08 0.04 0.00 0.04 0.20 0.19 0.19 0.15 0.15 0.16
A2_3 0.10 0.08 0.09 0.05 0.04 0.00 0.20 0.19 0.19 0.16 0.15 0.16
A4_1 0.20 0.20 0.21 0.20 0.20 0.20 0.00 0.06 0.05 0.10 0.11 0.11
A4_2 0.20 0.20 0.20 0.20 0.19 0.19 0.06 0.00 0.03 0.11 0.11 0.11
A4_3 0.20 0.20 0.20 0.20 0.19 0.19 0.05 0.03 0.00 0.11 0.10 0.12
A5_1 0.16 0.16 0.17 0.16 0.15 0.16 0.10 0.11 0.11 0.00 0.06 0.08
A5_2 0.17 0.16 0.16 0.15 0.15 0.15 0.11 0.11 0.10 0.06 0.00 0.06
A5_3 0.16 0.16 0.17 0.16 0.16 0.16 0.11 0.11 0.12 0.08 0.06 0.00
0.2
0.15
0.1
0.05
0

(B)


A1_1 A1_2 A1_3 A2_1 A2_2 A2_3 A4_1 A4_2 A4_3 A5_1 A5_2 A5_3
A1_1 0.00 0.00 0.01 0.01 0.01 0.01 0.04 0.04 0.04 0.03 0.03 0.03
A1_2 0.00 0.00 0.00 0.01 0.01 0.01 0.04 0.04 0.04 0.03 0.03 0.03
A1_3 0.01 0.00 0.00 0.01 0.01 0.01 0.04 0.04 0.04 0.03 0.03 0.03
A2_1 0.01 0.01 0.01 0.00 0.00 0.00 0.04 0.04 0.04 0.03 0.02 0.02
A2_2 0.01 0.01 0.01 0.00 0.00 0.00 0.04 0.04 0.04 0.02 0.02 0.02
A2_3 0.01 0.01 0.01 0.00 0.00 0.00 0.04 0.04 0.04 0.02 0.02 0.02
A4_1 0.04 0.04 0.04 0.04 0.04 0.04 0.00 0.00 0.00 0.01 0.01 0.01
A4_2 0.04 0.04 0.04 0.04 0.04 0.04 0.00 0.00 0.00 0.01 0.01 0.01
A4_3 0.04 0.04 0.04 0.04 0.04 0.04 0.00 0.00 0.00 0.01 0.01 0.01
A5_1 0.03 0.03 0.03 0.03 0.02 0.02 0.01 0.01 0.01 0.00 0.00 0.01
A5_2 0.03 0.03 0.03 0.02 0.02 0.02 0.01 0.01 0.01 0.00 0.00 0.00
A5_3 0.03 0.03 0.03 0.02 0.02 0.02 0.01 0.01 0.01 0.01 0.00 0.00
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0

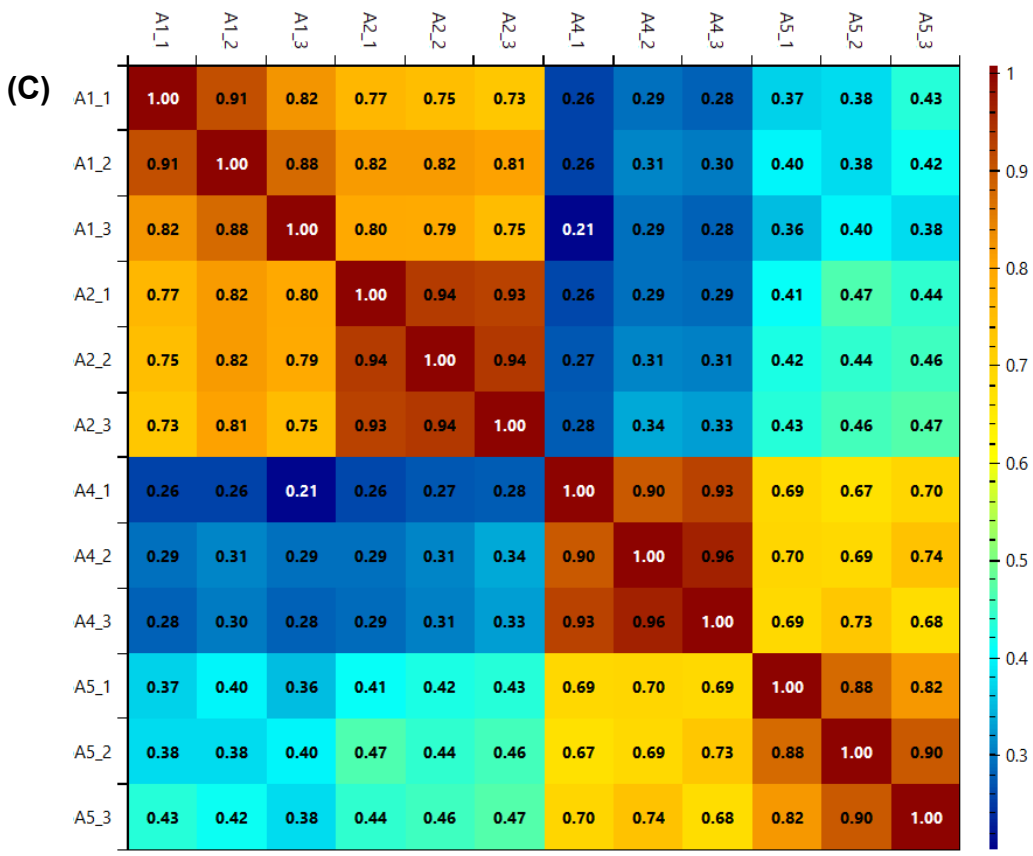


Figure 4-12 – Reproducibility heatmaps. The heatmaps generated using *k* (A), k^2 (B) and r^2 (C) as score metrics. It is possible to see that while the colour system changes, the information given is the same, the technical replicates (in groups of three in the order shown) are very similar, while some of the biological replicates are quite different.

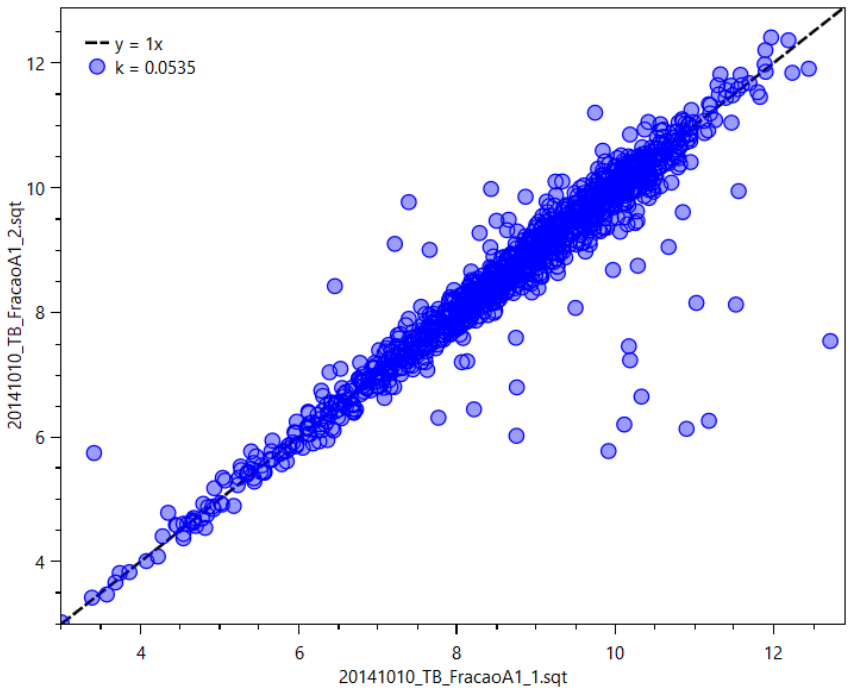


Figure 4-13 – Dot plot from reproducibility analysis. Dot plot comparing the XIC of common peptides of two different files, with the score shown in the left as the metric *k*, and a trendline for the equation $y = x$.

4.3 QUIN-XL

Having in mind the objective of more easily identifying regions of great structural change between proteins present in different conditions, we developed the software QUIN-XL, which applies a quantitative approach in order to acquire this information.

QUIN-XL can use data from previously identified cross-links separated by biological conditions to quantify them using XIC. These quantified cross-links will be clustered according to their quantitative profile, which will give insights of regions of variations or similarities between the conformers present in different states. These results can be easily browsed through, with a graphical interpretation provided, along with some options of exporting files so the user can continue the analysis in other computational tools.

4.3.1 Dataset processing

QUIN-XL makes use of cross-link identification files and raw MS files as input. The raw MS files come directly from the equipment used, while the identification file must be previously generated using SIM-XL [49]. In summary, SIM-XL works by creating theoretical spectra for possible cross-links from a given protein database and comparing them to experimental spectra, then assigning a score for the possible identifications. The software's output is shown on **Figure 4-14**.

After all the files from each biological condition are generated, they are loaded on the QUIN-XL software. There, the user will be able to label each file according to the biological condition based on how they should be separated for that particular experiment (**Figure 4-15**).

After all the files have been loaded, it is possible to set a range of the chromatography to be used for the normalization of the quantitative values (**Figure 4-16**). The normalization factor used is the Total Ion Current and consists of summing the intensities of all chromatographic peaks during the run. The ability to set a range is important as, usually, there are no peptides eluting in the first minutes of the chromatography, and the last minutes are reserved for washing the column. In those intervals, there are very few MS2 spectra being acquired and therefore, little to no identifications. To identify an optimal range, the user can check the TopN distribution on the RawVegetable software (explained in section **4.2.2**). This option enables identifying the frequency of MS2 acquisition throughout the chromatography and define an interval for the TIC range.

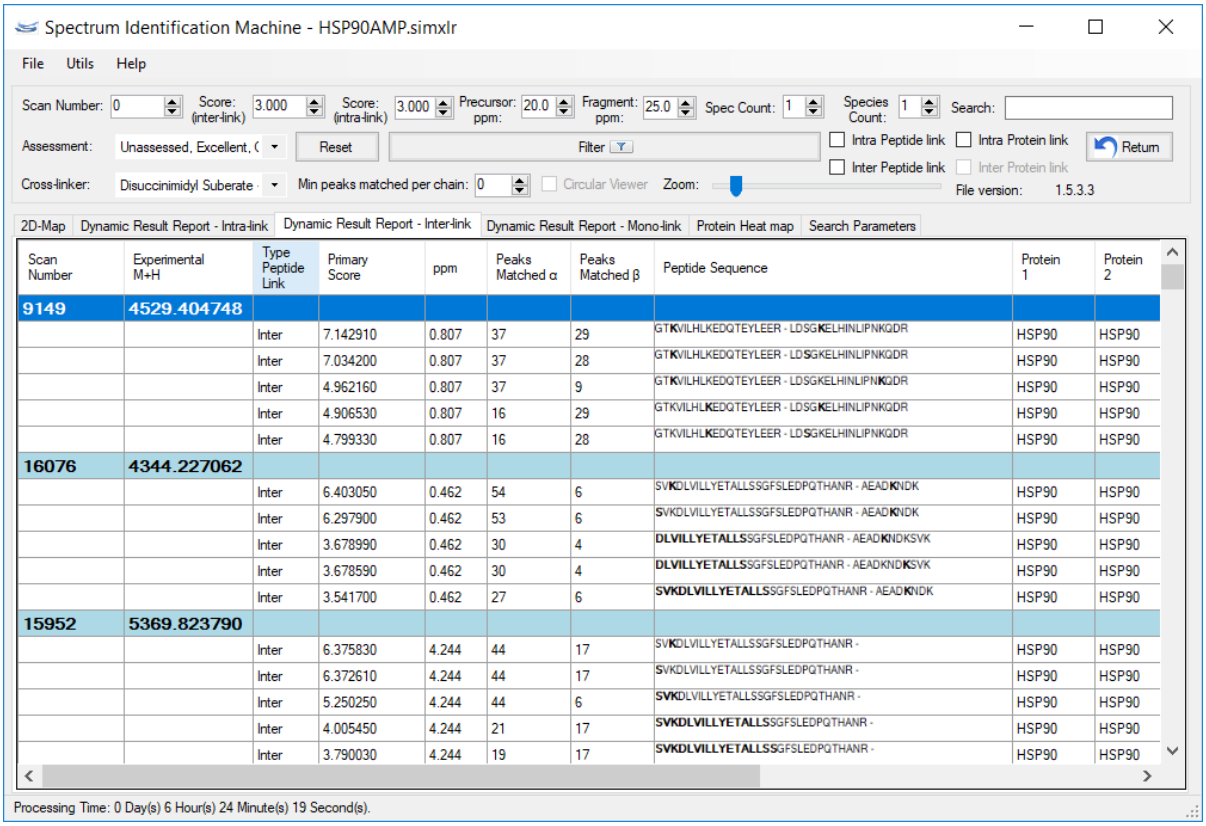


Figure 4-14 - SIM-XL result interface. Each blue row represents a spectrum, with all possible cross-links listed below. It is also possible to see the score of each identification, as well as the number of peaks that matched between theoretical and experimental spectrum.

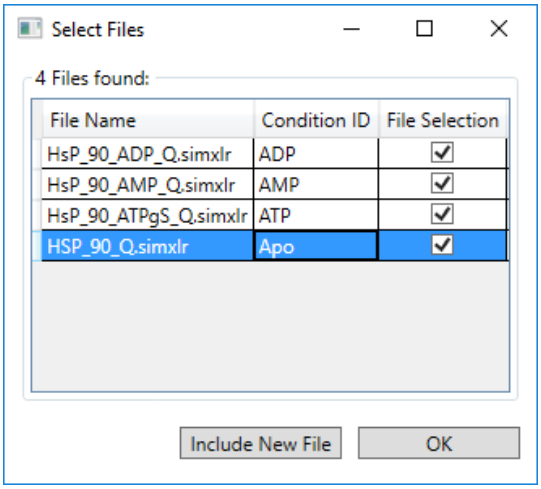


Figure 4-15 – File selection. The file selection interface allows the user to specify the SIM-XL files originating from different biological conditions and to assign a condition ID to each file. Note that multiple files can have the same ID.

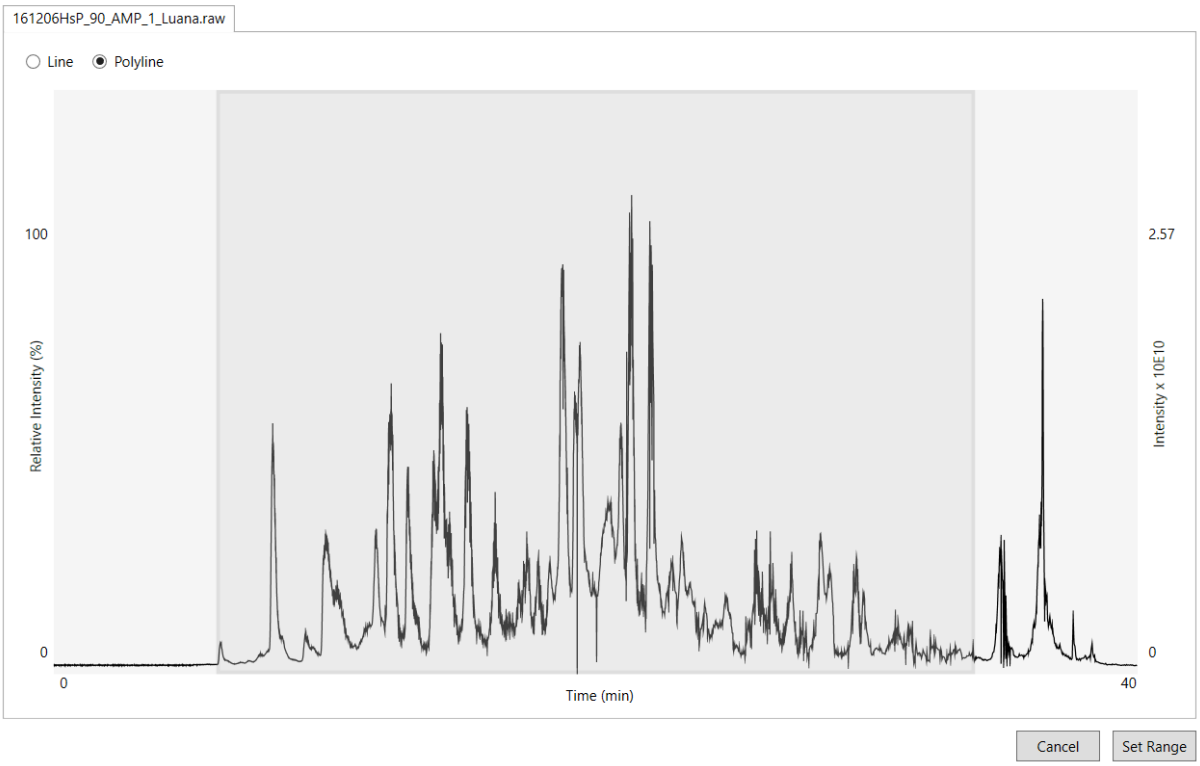


Figure 4-16 – Chromatographic range selection. Graphical interface showing the chromatographic profile of the run. Here, the user can set the range of intensities to be used for XIC normalization.

4.3.2 Quantification by XIC

Quantification values are calculated from the original mass spectrometry data as XICs, detailed in section **1.2.3**. The values can be displayed with or without normalization achieved by dividing the XIC value by the TIC determined as aforementioned.

Due to inaccuracies in ion intensity measurements and under-sampling, not all XIC curves are reliable. Trustworthy XICs usually have the approximate shape of a Gaussian curve, such as the ones in **Figure 4-17**. It is also possible to filter out curves with fewer than a certain number of points (e.g., 7), which might flag it as not representative enough of the sample. Either way, these unacceptable curves are not considered, and no quantitative value is assigned from them.

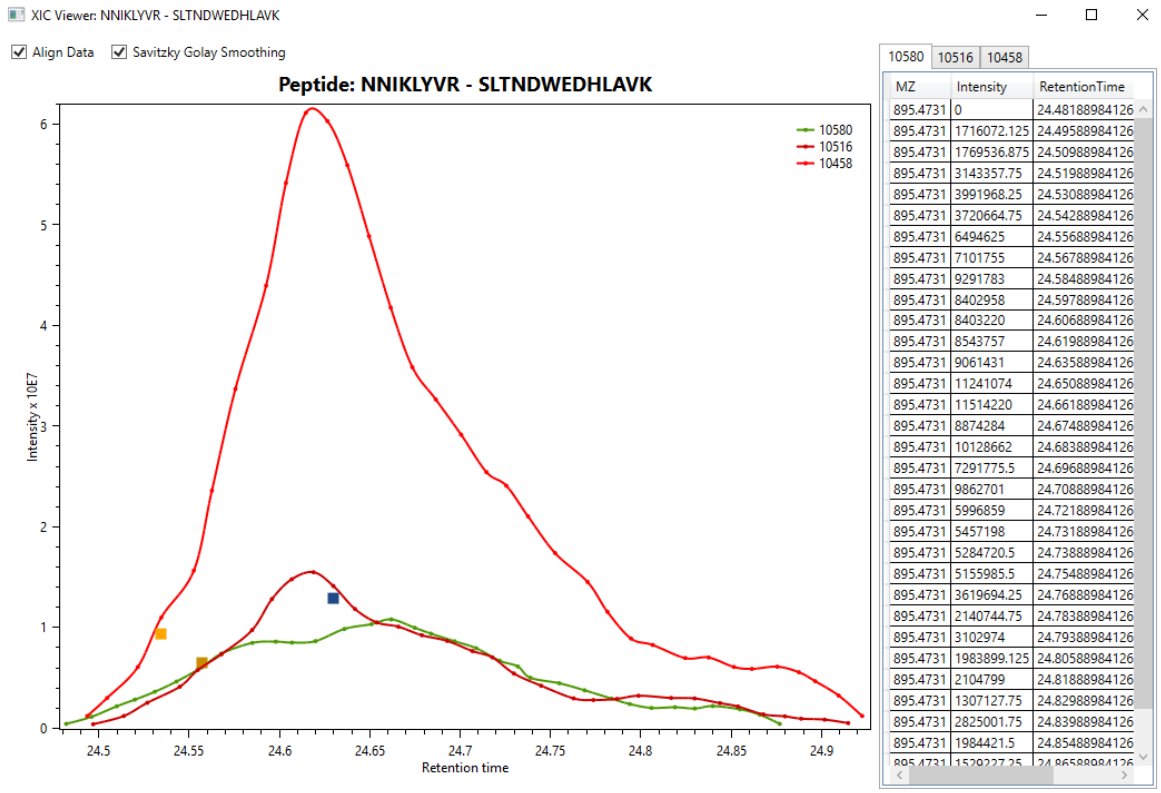


Figure 4-17 – XIC Viewer. Window showing all the XICs obtained for the selected cross-link at a given biological condition. The squares in the plot represent which MS2 scan was mapped to the identification of that XL.

Due to the time it takes to fill certain mass analyzers, a peak being tracked to extract XIC may sometimes vanish in a MS1 scan but shows up in the next one with a high ion current. This is a measurement error from the equipment that would lead, effectively, to an XIC curve being cut in two, as can be seen in **Figure 4-18 (A)**. To solve this, when a peak of interest disappears, the software looks at the next MS1 scan and if the peak shows up again, then the problematic MS1 is skipped. The software allows for only one such event to happen in an XIC curve, resulting in a curve shown in **Figure 4-18 (B)**.

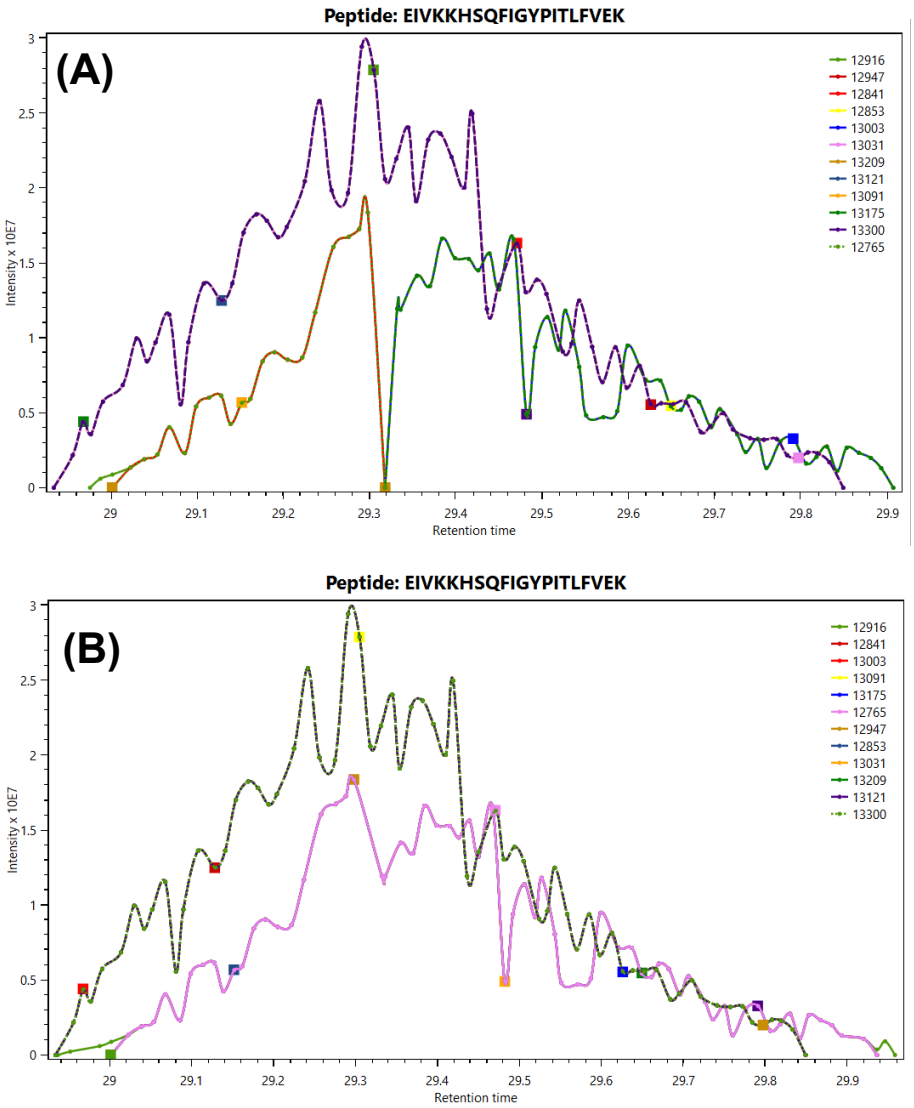


Figure 4-18 – XIC Correction. (A) XIC without tolerance, showing two different curves being formed (in orange and in green); (B) XIC with tolerance, now the problematic scan is skipped and a single curve (in pink) can be integrated.

Measurement inaccuracies and fluctuations from the equipment can generate spikes in XIC curves. As such, some users prefer to work with smoothed data (**Figure 4-17**). To achieve this we implemented an adapted Savitzky-Golay filter, first described in 1964 [61], then simplified by algebraic equations in 1978 [62]; briefly, this filter consists of trying to fit successive subsets of *m* points (from the total points *n* in XIC curve) to a polynomial to originate a new set of smoothed points, using the following equation:

$$y_j = \sum_{i=\frac{1-m}{2}}^{\frac{m-1}{2}} C_i y_{j+i} \qquad (8)$$

Where y_j is the new smoothed point being calculated; *j* is the number of the point from the total *n* XIC points; *m* is the subset number of points being used to fit a

polynomial (this has to be an odd number to count the current point in the center and a symmetrical window around it); i is the current point within the subset; y_{j+i} is the original value for y in the point with index $j + i$ and C_i is a convolution coefficient for the current point within the subset. This coefficient value can be taken from stablished tables or calculated according to the polynomial degree in the function used to fit the subset of points, using the equations described by Madden [62].

For example, applying the filter to the curve from **Figure 4-18 (B)** would result in **Figure 4-19**, which has many of its spikes removed, possibly increasing precision while not distorting the original data.

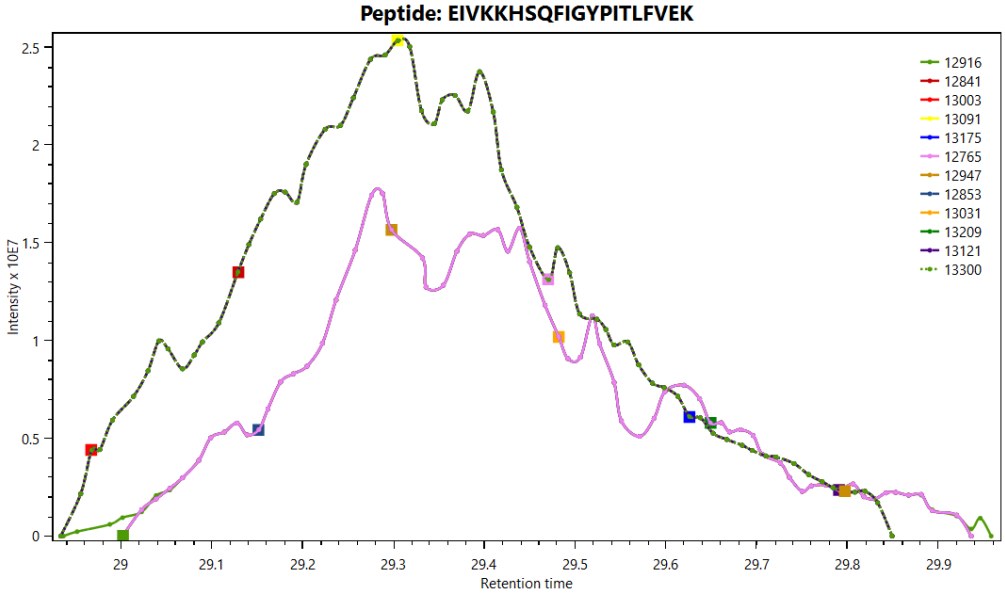


Figure 4-19 – XIC smoothing. XIC curve smoothed using our implementation of the Savitzky-Golay filter.

After all these filters are applied on the XIC curves, either the one with the best area to number of points ratio is chosen to represent the quantitative value of that identification, or an average of all the curves extracted can be calculated.

When all the data is finished processing, the software lists all identified cross-links and their corresponding quantitative (XIC) values, for each biological condition, in a dynamic report. The software allows the user to single out cross-links exclusive to a condition based on their quantitative value, and filter cross-links by intralinks (cross-links in the same peptide), interlinks (between two different peptides), scores, cross-linker used or simply search for a given sequence. It is also possible to show the XIC value of a residue pair, that is, group multiple peptides according to the position of the cross-link relative to the protein. Since a single residue pair can be identified in different

peptide links (usually with a few amino acids more or a modification in the same peptide), it is interesting to look at the residue pair, instead of the whole sequence.

The GUI showing all these results and filters can be seen on **Figure 4-20**.

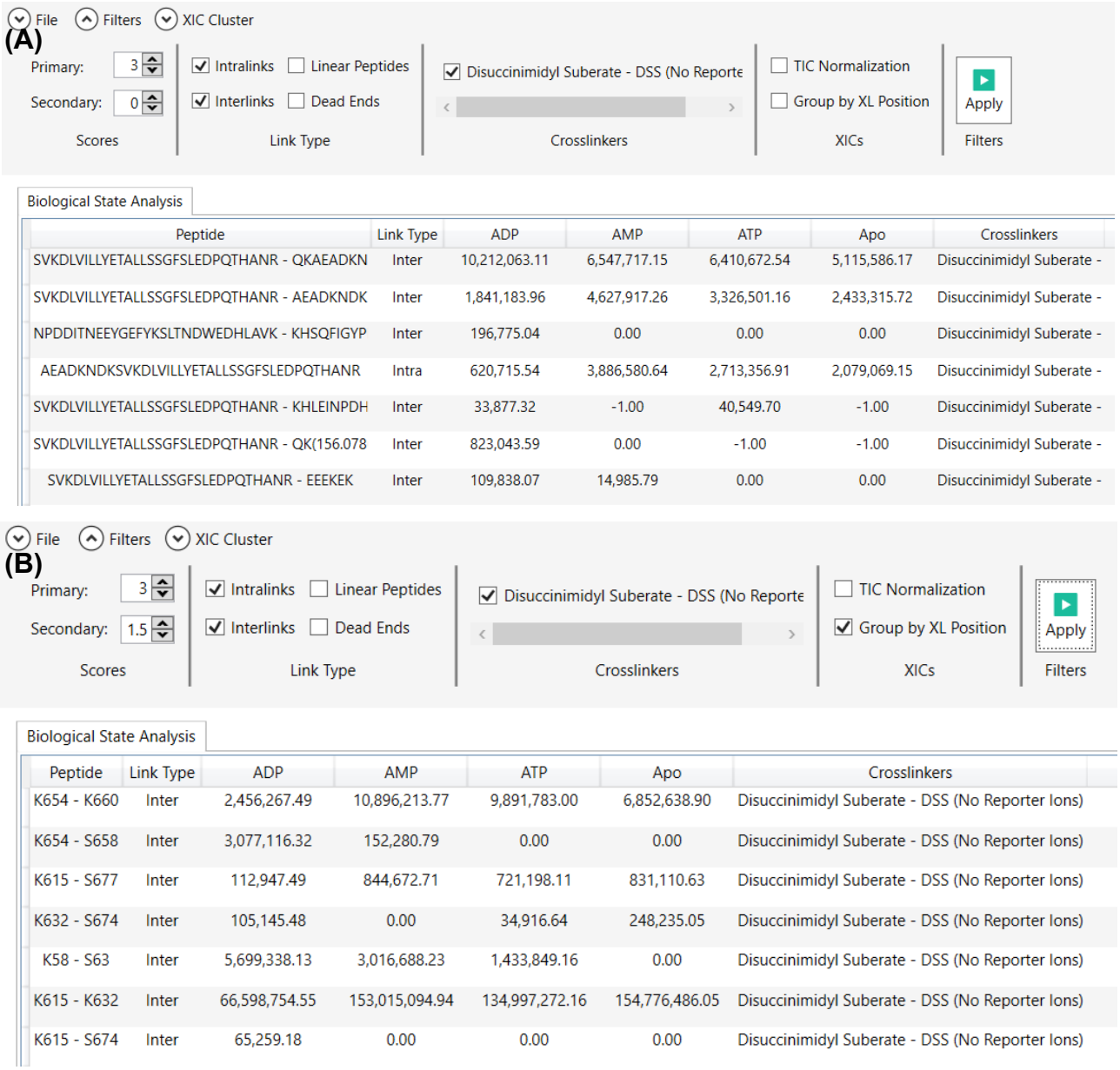
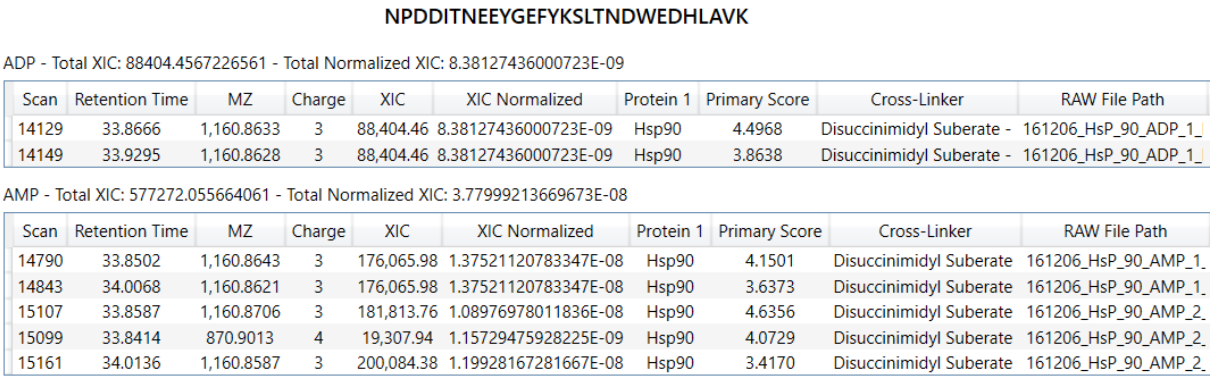


Figure 4-20 – QUIN's main interface. (A) QUIN-XL main window: column 1 contains the peptide's sequence from the identified cross-links; column 2 shows the link type as either intralink or interlink; columns 3 to 6 are the total XIC values for each of the assigned biological states, four in this case (a value of zero means that the cross-link was not identified in that specific condition; -1 means that the cross-link exists but the software could not determine a quantitative value, probably because there were fewer than three points in the XIC curve); column 7 lists the cross-linkers used in the experiment. It is also possible to see the filters that can be applied to the data on the top of the window. (B) Same screen but showing the cross-linked peptides grouped by the position of the amino acids where the cross-link appeared.

Double clicking on the sequence of a cross-link entry opens a window displaying complementary information such as from which proteins the peptides map to, all the

scans that were identified with that cross-link, along with all the XIC values that contributed for the total XIC displayed on the Main Window (**Figure 4-21**).



NPDDITNEEYGEFYKSLTNDWEDHLAVK

ADP - Total XIC: 88404.4567226561 - Total Normalized XIC: 8.38127436000723E-09

Scan	Retention Time	MZ	Charge	XIC	XIC Normalized	Protein 1	Primary Score	Cross-Linker	RAW File Path
14129	33.8666	1,160.8633	3	88,404.46	8.38127436000723E-09	Hsp90	4.4968	Disuccinimidyl Suberate -	161206_HsP_90_ADP_1_
14149	33.9295	1,160.8628	3	88,404.46	8.38127436000723E-09	Hsp90	3.8638	Disuccinimidyl Suberate -	161206_HsP_90_ADP_1_

AMP - Total XIC: 577272.055664061 - Total Normalized XIC: 3.77999213669673E-08

Scan	Retention Time	MZ	Charge	XIC	XIC Normalized	Protein 1	Primary Score	Cross-Linker	RAW File Path
14790	33.8502	1,160.8643	3	176,065.98	1.37521120783347E-08	Hsp90	4.1501	Disuccinimidyl Suberate	161206_HsP_90_AMP_1_
14843	34.0068	1,160.8621	3	176,065.98	1.37521120783347E-08	Hsp90	3.6373	Disuccinimidyl Suberate	161206_HsP_90_AMP_1_
15107	33.8587	1,160.8706	3	181,813.76	1.08976978011836E-08	Hsp90	4.6356	Disuccinimidyl Suberate	161206_HsP_90_AMP_2_
15099	33.8414	870.9013	4	19,307.94	1.15729475928225E-09	Hsp90	4.0729	Disuccinimidyl Suberate	161206_HsP_90_AMP_2_
15161	34.0136	1,160.8587	3	200,084.38	1.19928167281667E-08	Hsp90	3.4170	Disuccinimidyl Suberate	161206_HsP_90_AMP_2_

Figure 4-21 – Detailed peptide's information screen. Window showing a deeper comparison between biological states for a single selected cross-link. On this interface, it is possible to see all the XICs that contributed for the total XIC displayed. In the case of interaction between two different protein, it is also possible to see from which protein each peptide came from.

Double clicking on the XIC value for a given cross-link on the Main Window will open a window showing all the quantification curves obtained for that XL in the selected biological condition (**Figure 4-17**). Usually there are quite a few XICs for a single XL, which can come from different charge states of the same molecule or from an isotope.

4.3.3 XIC Clustering

In what follows, the software will then attempt to cluster molecules with similar quantitative profiles. To achieve this, the software first encodes each cross-link quantitative profile as an input vector; each argument of the vector corresponds to the XIC area for a given cross-link for a given condition. The input vectors are normalized by calculating the Euclidean norm of the vector and then dividing all the arguments by the norm, so that the new norm is 1, as shown below (where XL1 is the vector for a single cross-link):

$$XL1 = [XIC_{condition1}, \quad XIC_{condition2}, \quad XIC_{condition3}] \quad (9)$$

$$NormXL1 = \sqrt{XIC_{condition1}{}^2 + XIC_{condition2}{}^2 + XIC_{condition3}{}^2} \quad (10)$$

$$XL1_{normalized} = \left[\frac{XIC_{condition1}}{NormXL1}, \frac{XIC_{condition2}}{NormXL1}, \frac{XIC_{condition3}}{NormXL1}\right] \quad (11)$$

Now, if a new norm of the resulting vector is calculated, the result will be 1, as shown below:

$$NewNormXL1 = \sqrt{\left(\frac{XIC_{condition1}}{NormXL1}\right)^2 + \left(\frac{XIC_{condition2}}{NormXL1}\right)^2 + \left(\frac{XIC_{condition3}}{NormXL1}\right)^2} \quad (12)$$

$$NewNormXL1 = \sqrt{\frac{{XIC_{condition1}}^2 + {XIC_{condition2}}^2 + {XIC_{condition3}}^2}{NormXL1^2}} \quad (13)$$

$$NewNormXL1 = \frac{\sqrt{{XIC_{condition1}}^2 + {XIC_{condition2}}^2 + {XIC_{condition3}}^2}}{NormXL1} \quad (14)$$

$$NewNormXL1 = \frac{\sqrt{{XIC_{condition1}}^2 + {XIC_{condition2}}^2 + {XIC_{condition3}}^2}}{\sqrt{{XIC_{condition1}}^2 + {XIC_{condition2}}^2 + {XIC_{condition3}}^2}} \quad (15)$$

$$NewNormXL1 = 1 \quad (16)$$

This is done for all cross-link vectors; if the vector is not normalized, the clustering will happen according to XIC intensities, while what we want is a clustering by quantitative patterns.

The clustering algorithm chosen to group cross-link identifications with similar quantitative profile is the *k*-means algorithm [63]. In brief, this method works by plotting the vectors in a space of *n*-dimensions, where each dimension is a biological condition. For a given number of clusters (*k*), *k* random points are set in the *n*-dimensional space to be the centroids of the clusters, as shown in **Figure 4-22**. Each data-point is assigned to a cluster, according to its closest centroid using an Euclidean distance metric (Equation (7)). A new set of centroids is then calculated as the average of all points in the new clusters. This process of adjusting centroid positions is repeated until the centroids stabilize in a position.

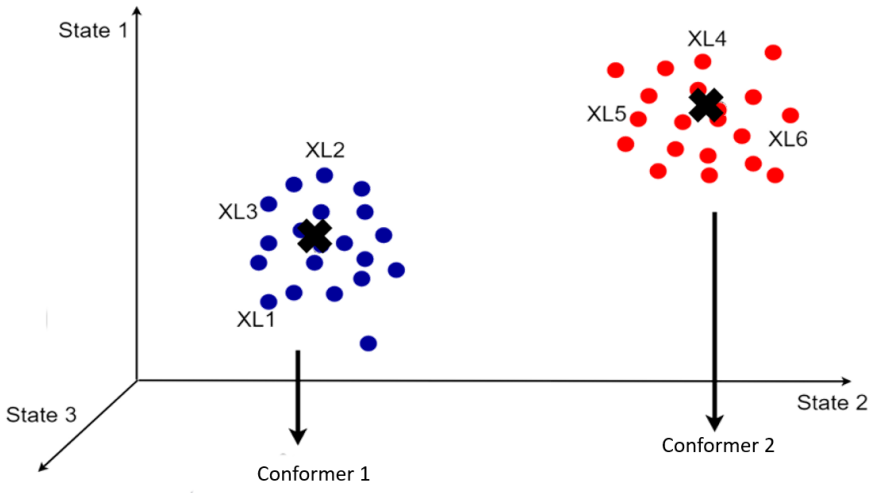


Figure 4-22 – XL Clustering. Representation of a set of cross-links (XLs) from three biological states (so a three-dimensional space) being clustered into two groups (in blue and red), so *k* = *2*. Centroids

are shown as black crosses. As different conformers should have different quantitative profiles, each cluster should represent a different conformer or region of the protein being studied.

The input for traditional *k*-means implementations requires a specific number of clusters, which is not always obvious at the outset of a problem. Therefore, we implemented a Silhouette-scoring based approach for inferring an optimal number of clusters [64]. Here, the *k*-means algorithm is executed several times, at increasing number of clusters. A total score is attributed for each of these runs; defined as the sum of individual Silhouette scores for each cluster member. The individual scores are equal to the average of the Euclidean distances to all vectors within its own cluster and in the nearest cluster. The equation used for calculating the Silhouette score is the following:

$$S = \sum_{i=0}^{n} \frac{b_i - a_i}{\max\{a_i, b_i\}} \qquad (17)$$

Where *S* is the score; *n* is the total number of cross-links identified and being clustered; *i* is the index of the current XL point being analysed; *a* is the average of the Euclidean distance of the point of XL_i to all XLs in the same cluster; *b* is the average of the Euclidean distance of the point of XL_i to all XLs in the nearest cluster; $max\{a_i, b_i\}$ means that the highest number between *a* and *b* will be used as divisor.

The optimal number of clusters will be the one leading to the highest total score (e.g. run A had the number of clusters set as 5 and a total score of 100 and run B had the number of clusters set as 7 and a total score of 80; the best run will be A and the optimal number of clusters is 5). The graphical result of the Silhouette method can be seen on **Figure 4-23**.

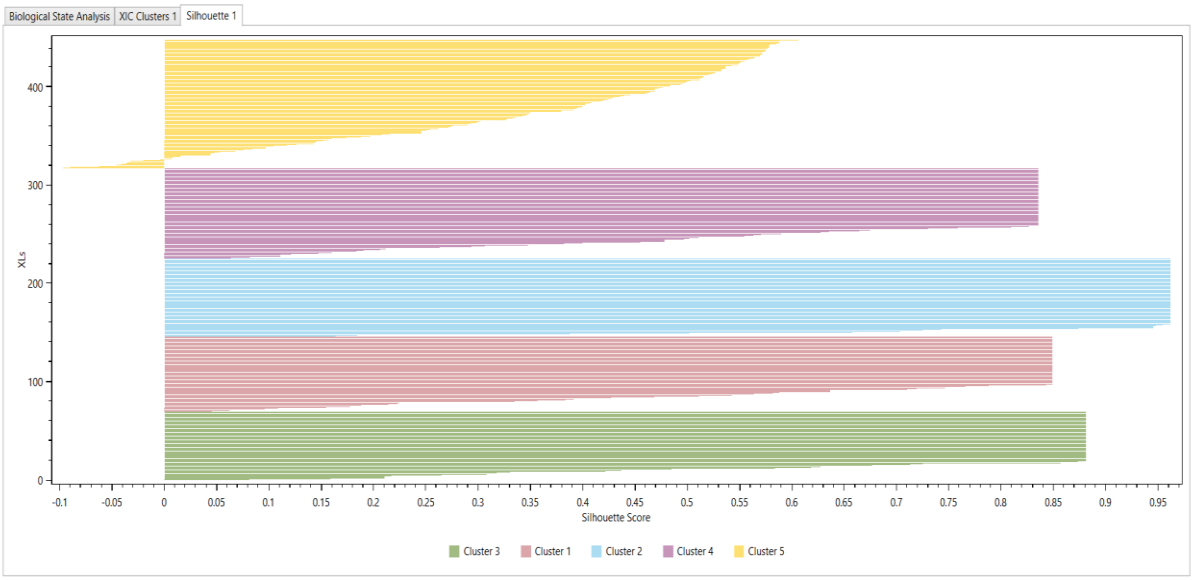


Figure 4-23 – Silhouette Plot. Graphical result of the best run determined by the Silhouette method. Here, each row represents the silhouette score of a single cross-link, grouped in different colours according to assigned cluster. If the score is close to 1, the XL is in the correct cluster; if it is close to 0, the XL is probably in a grey-zone between two clusters; and if it is close to -1, then the XL is probably in the incorrect cluster and should be revaluated or manually reassigned.

After the number of clusters has been set and the cross-links have been assigned to a cluster, they will be shown on the screen in the form of linear plots, where each plot is a cluster and each line is a cross-link, with the *y*-axis being the normalized XIC area according to each biological condition represented on the *x*-axis (**Figure 4-24**).

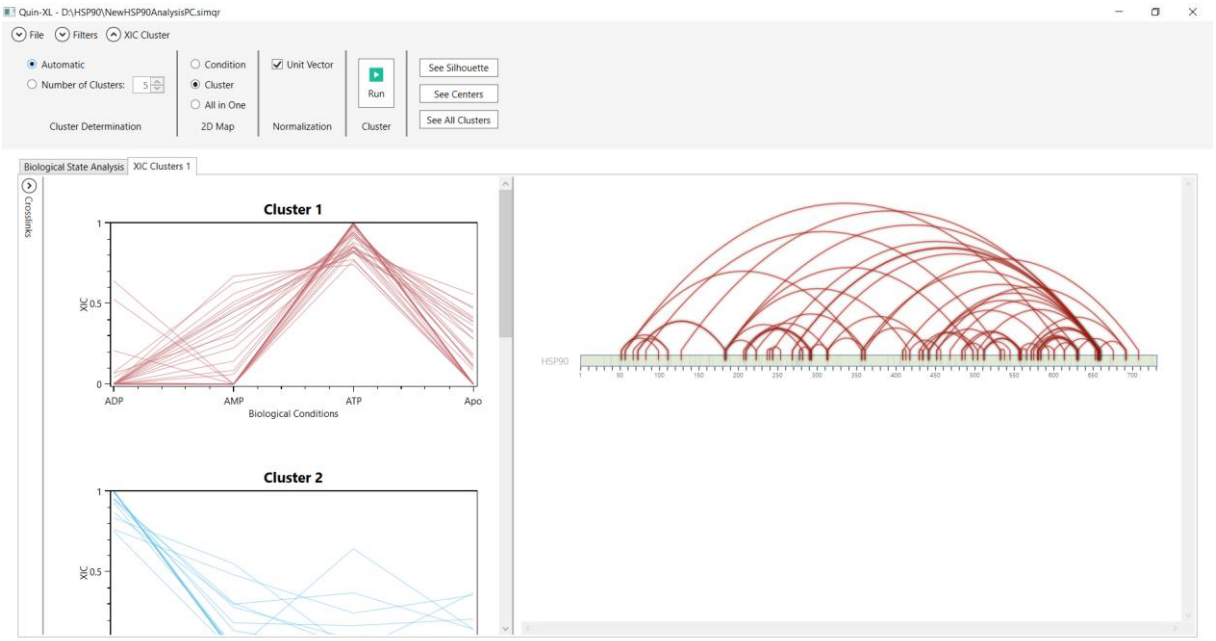


Figure 4-24 – QUIN's clustering result screen. Screenshot of the resulting screen showing all quantitative profile clusters generated on the left side, where each line in a plot is a different cross-link identified; and a 2D representation of the protein and its XL on the right side. The menu "Cross-links" on the left side (collapsed) contains a table with all XLs; clicking on one of them highlights the corresponding curve on one of the cluster plots.

The k-means clustering algorithm is not deterministic, meaning different runs can lead to slightly different results, and even different optimal number of clusters (*k*). Furthermore, automatically obtained clusters might prove to be too populated to provide easy to understand insights on the structure. With this in mind, we implemented functionalities for selecting a single cluster, and have the algorithm recluster only that specific set of cross-links (**Figure 4-25**). The user can then reject or accept the new clusters, in which case they will show alongside the previous ones on the main window.

Finally, the user can also manually change a cross-link from one cluster to another by simply changing its label on the table that lists them all.

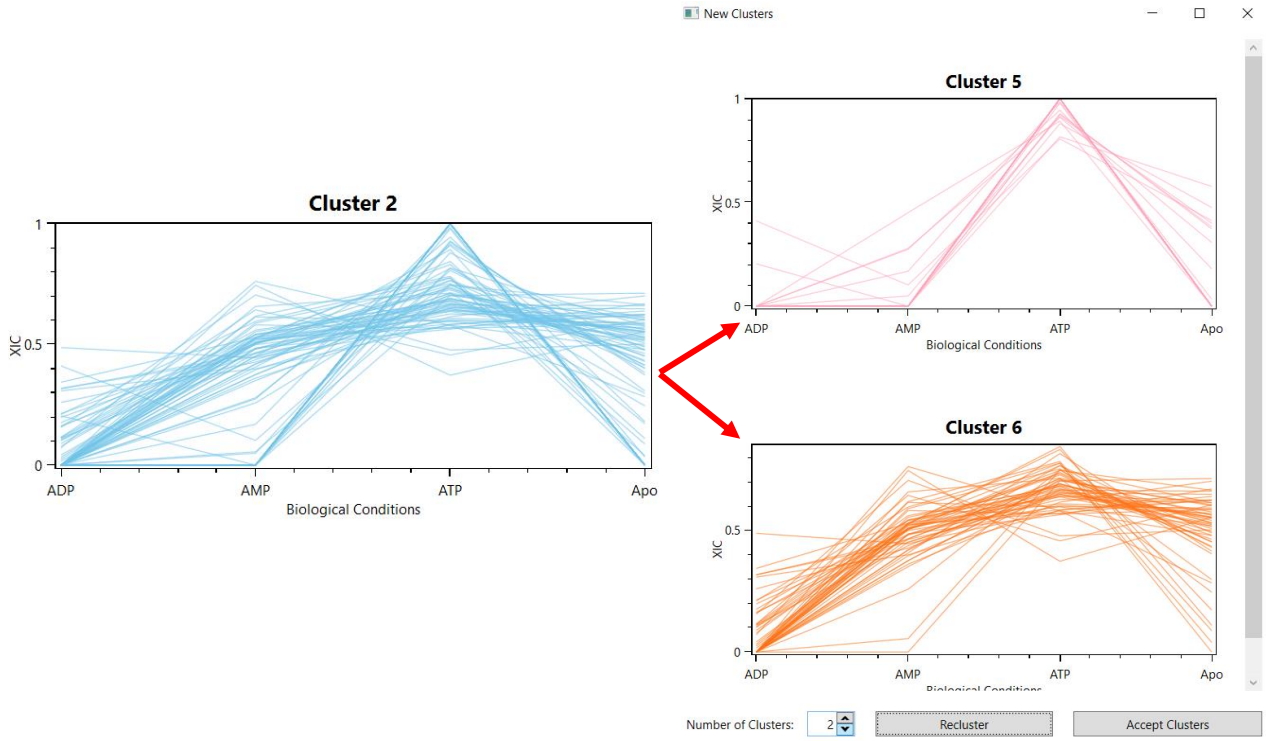


Figure 4-25 – Reclustering example. Example showing the two new clusters generated from reclustering Cluster 2 (in blue). Those can be accepted, rejected or reclustered with different number of clusters.

4.3.4 Protein Mapping

Once the quantitative-profile clusters are generated, it is possible to map individual cross-links within a cluster to the protein and have a better view of their position. This can be done by double clicking on one of the cluster plots, which shows a linear representation of the protein, called a 2D-Map (**Figure 4-24**). This map is the same one seen on the SIM-XL software. Here the protein is shown as a bar, and all the cross-links for the selected cluster as red arches (**Figure 4-26**). This visualization strategy

makes it easier for the user to spot differences between proteins structures, by assessing which regions are more enriched with cross-links and choose which cross-links to select for a possible structural modelling.

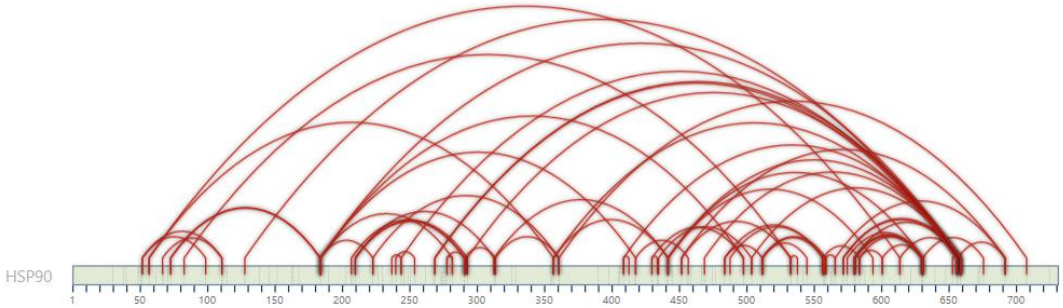


Figure 4-26 – Protein's 2D-Map. A 2D-Map of the HSP90 protein (represented as the light green bar) with all the cross-links identified for selected cluster (each cross-link represented as a red arch). The numbers under the bar are amino acids positions.

Aside from looking at the linear representation of the protein, the user can compare the XLs in each cluster with a protein structure by exporting a file in the format that the software Topolink [65] reads. Topolink measures the topological distances between possible cross-link sites in protein structure models (crystallographic or computational), which must be in the .pdb format (from the Protein Data Bank). These distances are then compared with the maximum distance for the cross-linker molecule to validate the cross-links found. It is also possible to eventually use the information from the clusters to improve computational models for protein structures [66].

Cross-links in each cluster can be exported as PyMOL scripts (.pml file), which will show the XLs as Euclidean distances when it is run on a protein structure (.pdb file), as in **Figure 4-27**. This enables seamlessly verifying a cross-link cluster in the 3D structure, and thus where most of the XLs are coming from, and therefore providing insights on variable structural regions between conformers.

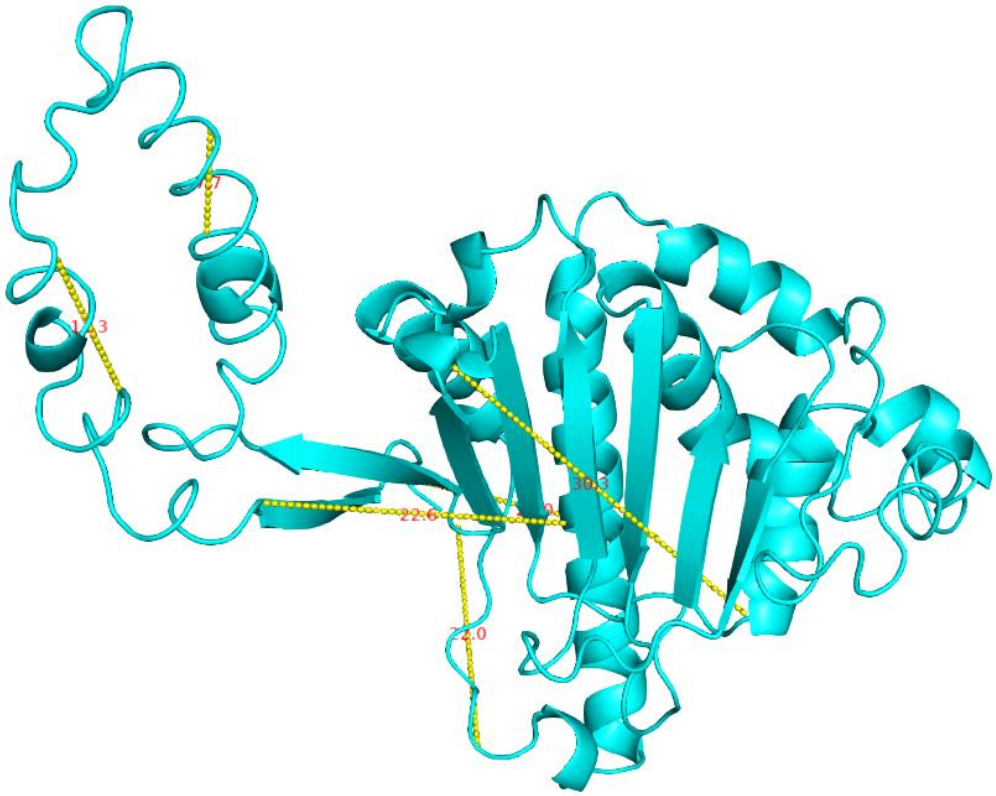


Figure 4-27 – Protein's 3D model. Representation of a protein model of the NTD of HSP90 in PyMOL, shown in cyan. In yellow are the cross-links represented as Euclidean distances (value in Angstroms shown in red).

5. EXPERIMENTAL VALIDATION AND DISCUSSION

For the experimental validation of the developed methodology, the molecular chaperone HSP90 was chosen as the target. HSP90 was chosen because it has a known conformational cycle, with a defined crystal structure in most of these conditions, except in the presence of AMP, thus becoming a valuable target for the experimental validation. The target sample was separated in four biological conditions: with ATP; with ADP; with AMP and without any nucleotides.

In what follows we will describe the data acquisition, chromatography QC, and finally the quantitative-profile based conformer analysis performed by QUIN-XL, and the insights the software was able to provide on the biochemical characteristics of the HSP90 protein in accordance to its available structures.

5.1 Data acquisition

All the experimental procedure for protein purification, cross-linker reaction, and mass spectrometry data acquisition was performed by the Dalton Mass Spectrometry Laboratory at UNICAMP, which is led by Dr. Fabio Gozzo.

The HSP90 protein used for the experiments was the human one and was analysed in a purified solution, with the nucleotides being added separately afterwards for the formation of complexes. The nucleotides were ADP, AMP and ATP-γ-S (referenced in the text as ATP), which is a more stable analogue of ATP and will not hydrolyse immediately. The cross-linker used was DSS in a molar excess of 100 times relative to the protein concentration. The cross-linked proteins were digested using trypsin.

Mass spectrometry data for HSP90 was generated in the following conditions: a) ATP-bound, b) ADP-bound, c) AMP-bound, and d) nucleotide-free (Apo); all using an Orbitrap Fusion Lumos Mass Spectrometer (ThermoFisher Scientific). The equipment was set to acquire up to eight MS2 spectra per duty cycle. For each condition, two technical replicates were produced in the mass spectrometer, creating a total of eight spectra files to be used in the identification. The data from each condition was analysed separately using the SIM-XL software with to the following parameters:

- Cross-linker: DSS;
- Precursor ppm (MS1): 10;
- Fragment ppm (MS2): 20;
- Modifications: Carbamidomethylation of Cystein.

5.2 Chromatography analysis with RawVegetable

The eight RAW files generated were analysed with RawVegetable in order to check the quality of the experiment. Information for each run can be seen on **Table 5-1**.

Table 5-1 - General information on the MS run for the HSP90 experiment. This table shows the four biological conditions (Apo, ADP-bound, AMP-bound and ATP-bound) and their respective MS1 and MS2 total spectrum count, average time for a duty cycle and the full time of the chromatography.

Condition	MS1 Spectra	MS2 Spectra	Duty Cycle Average Time (min)	Chromatography Full Time (min)
Apo_1	4735	12958	0.0084	39.9800
Apo_2	4722	12583	0.0084	39.9887
ADP_1	4312	11915	0.0092	39.9906
ADP_2	4096	9481	0.0097	39.9981
AMP_1	4692	12261	0.0085	39.9994
AMP_2	4761	12478	0.0084	39.9984
ATP_1	4742	6028	0.0084	39.9974
ATP_2	4731	12579	0.0084	39.9953

Most of the experiments had similar characteristics to each other, as well as chromatographic runs with similar profiles (**Figure 5-1**).

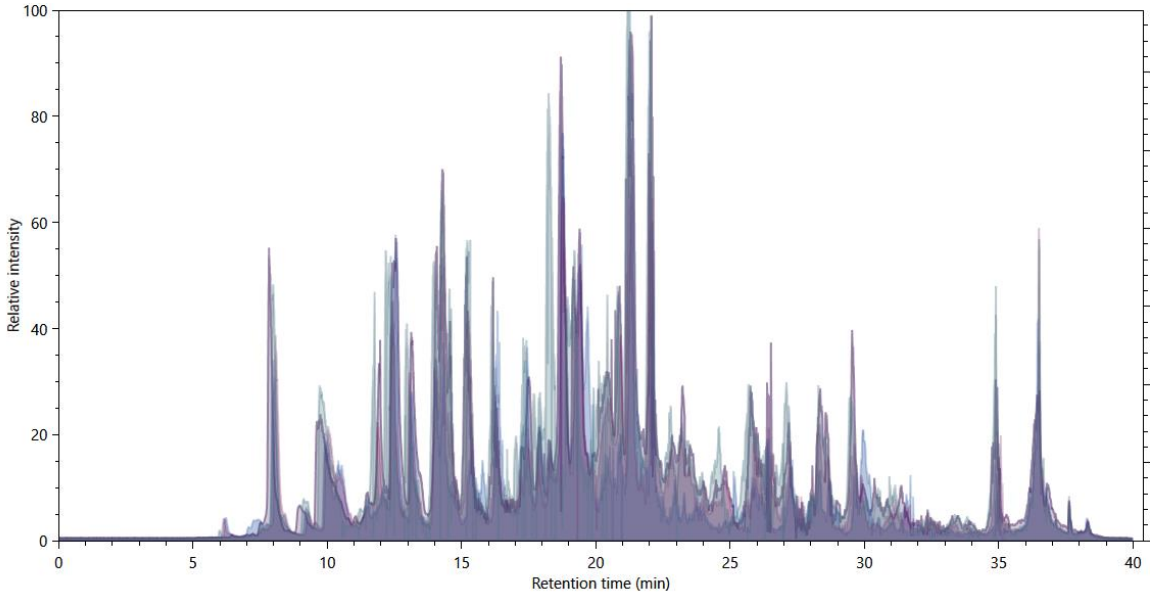


Figure 5-1 – Chromatographies from HSP90 experiment. Graph showing four overlapping chromatographies of the different conditions for the HSP90 experiment. Overall, the chromatographies are very similar to each other, which is to be expected, as the changes in the protein are not so drastic.

While most of the chromatographic runs had a nice profile, one of the replicates from the ADP experiment (ADP_2) and one from the ATP experiment (ATP_1) presented differences from other technical replicates and from what was expected given that the experiments are not so different between themselves. This shows an issue in data-generation leading to a lower number of MS2 spectra in comparison to the other runs, which can be confirmed by looking at their TopN density estimation (**Figure 5-2**).

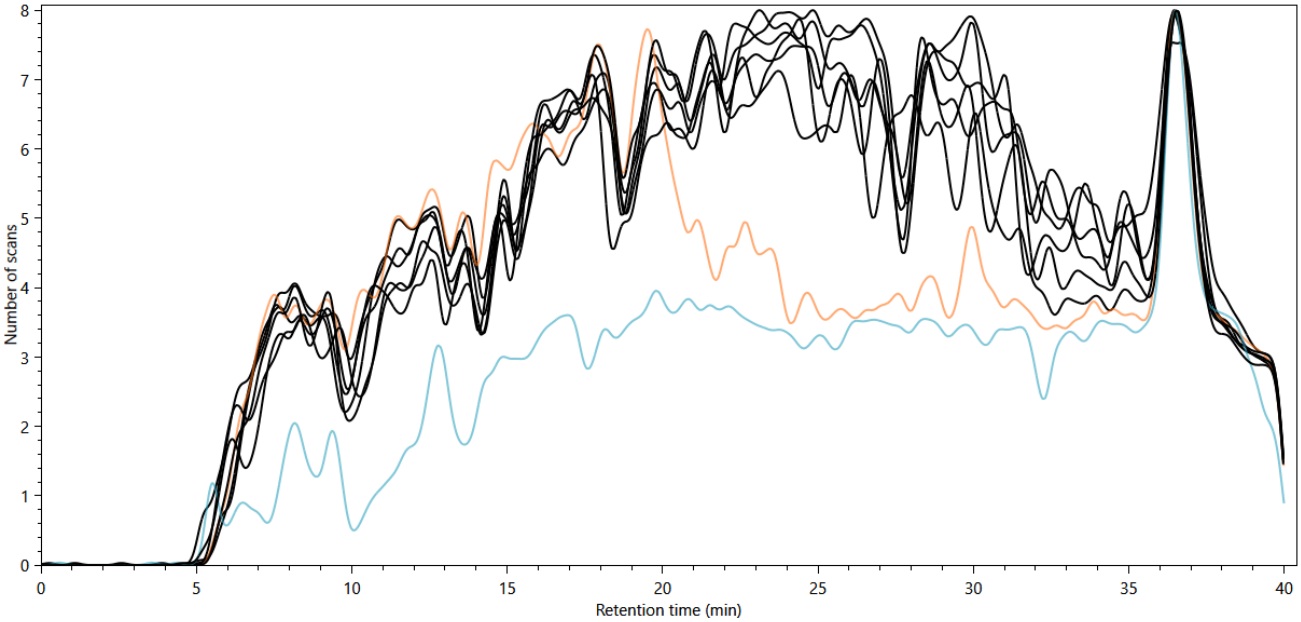


Figure 5-2 – TopN distributions from HSP90 experiment. The TopN distribution plot for all the HSP90 files; in orange is one technical replicate from ADP; in blue is one replicate from the ATP condition; and in black are all the other files. It is possible to see that while most of the runs show a similar distribution, the two highlighted replicates vary a lot.

The TopN distribution shows that the ADP_2 replicate starts off with a very similar profile, but then for half of the chromatography the density of MS2 events is much lower. The ATP_1 replicate seems to follow the same profile as most of the others but with much lower density throughout. While these two replicates are somewhat problematic when compared to the others, they can still be used for the identification of cross-links, especially since they have another replicate to back the experiment, but they might provide much less information than otherwise, as can be seen on **Figure 5-3 (A)**, which is the charges 3+ and 4+ chromatogram for ATP_1. When compared with the same charged chromatograms for ATP_2 (**Figure 5-3 (B)**), there were fewer identifications for ATP_1. Mostly this means that identifications coming from ATP_2 without confirmation from ATP_1 should be taken in carefully, as there is no way to guarantee if it was a misidentification or the XL did not show on ATP_1 because of the problems in the chromatography.

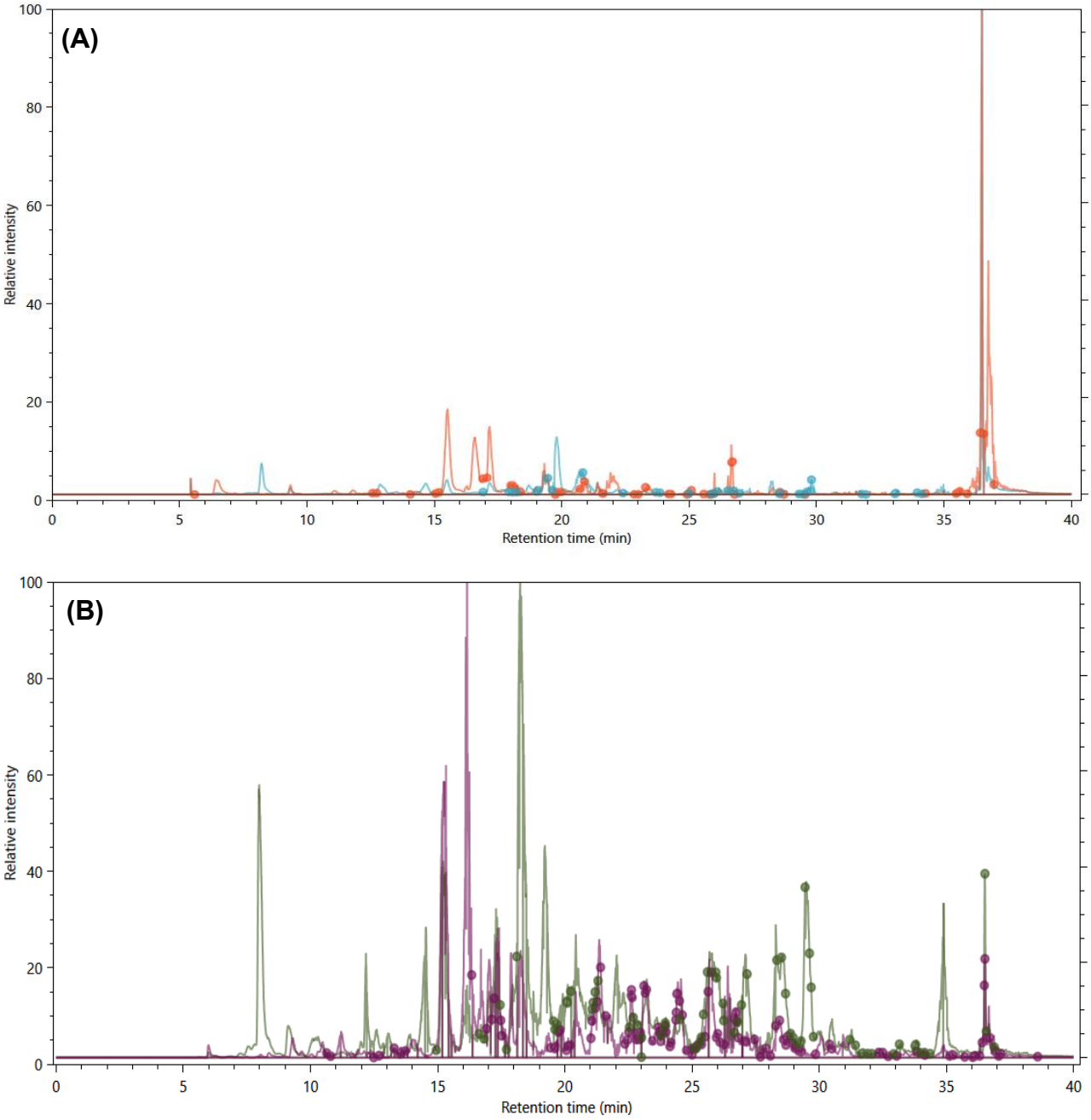


Figure 5-3 – Charged chromatograms for ATP sample. (A) Normalized chromatograms for charges 3+ (blue) and 4+ (orange) of the ATP_1 file, with the dots representing identifications events. (B) The normalized chromatograms for charges 3+ (green) and 4+ (purple) of the ATP_2 file. It is possible to see that the density of identifications is much lower in the ATP_1 file (A) when compared to ATP_2 (B).

The other files show a profile and XL identification density similar to ATP_2 (**Figure 5-3 (B)**). We can see that most of the identified XLs elute in the second half of the chromatography, from 20 minutes onwards. This indicates that, should this experiment be repeated, that region would be a prime target for gradient optimization to increase the number of informative MS2 scans.

The XL-Artifact algorithm was employed to check for the presence of non-covalently bound peptides that might lead to misidentifications. Of all eight files, the

only ones that showed a possible artifact with a score higher than 2.5 were ADP_1 and Apo_1. ADP_1 showed only one possible artifact, which was the cross-link between Lys-431 and Lys-499. Apo_1 showed another two, which are the same XL (between Lys-74 and Lys-185) with two different charges (5+ and 6+), as shown in **Figure 5-4**.

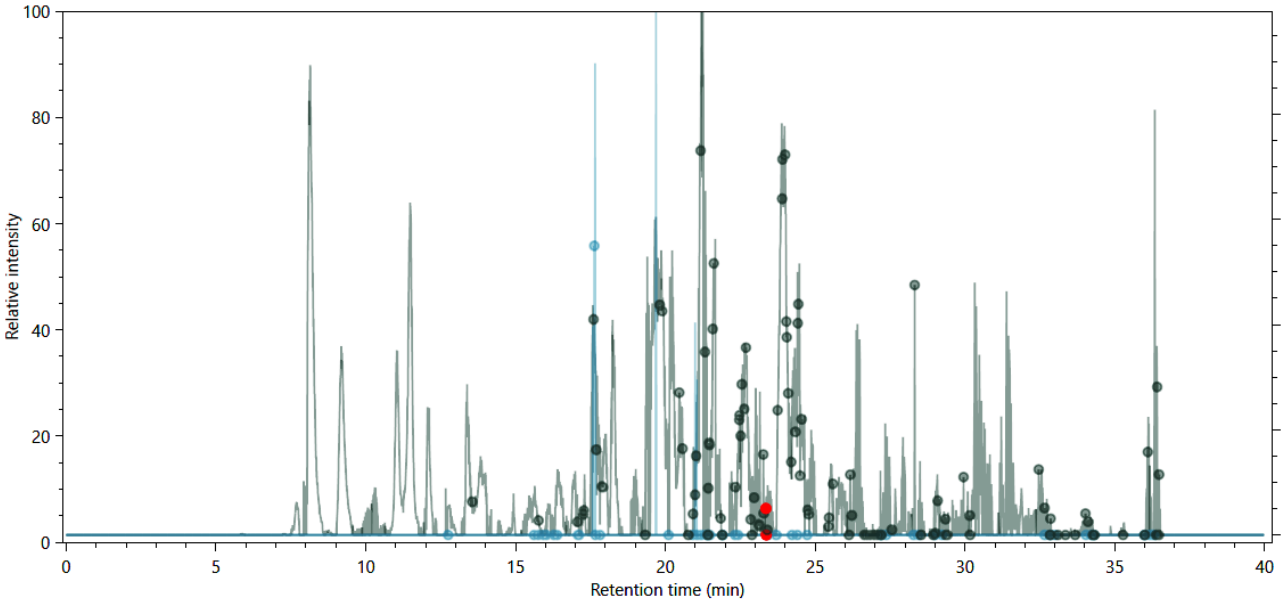


Figure 5-4 – Charged chromatograms for Apo sample. Normalized chromatograms for charges 5+ (green) and 6+ (blue) of the Apo_1 file. Highlighted as red dots are the two possible artifacts found, which are for the same XL but with different charges.

Overall, the chromatography runs presented the expected characteristics, as far as number of spectra, identifications, and reproducibility, with only the XLs from the ATP experiments requiring some attention post-analysis.

5.3 HSP90 clustering and validation

After the cross-links were identified using SIM-XL, they were loaded into QUIN-XL. The biological conditions were set in the same way as previously described (ADP-, AMP- and ATP-bound and Apo state), and each XL was quantified using the XIC approach. After the quantification, the XLs were filtered to have a minimum primary score of 3.0, secondary score of 2.0, show only inter and intralinks, a minimum of 7 points in the XIC curve and were grouped according to the cross-link site, not by the peptide, which resulted in a total of 107 cross-links to be clustered.

The clustering algorithm automatically performed tests for optimal number of clusters (*k*) from 2 to 15 and, with the Silhouette scoring system, concluded that the optimum value of *k* was five (**Figure 5-5**). Looking at the silhouette plot for the five clusters (**Figure 5-6**), we see that all XLs have positive scores, which means that most are within the right cluster and only a few in grey zones between clusters.

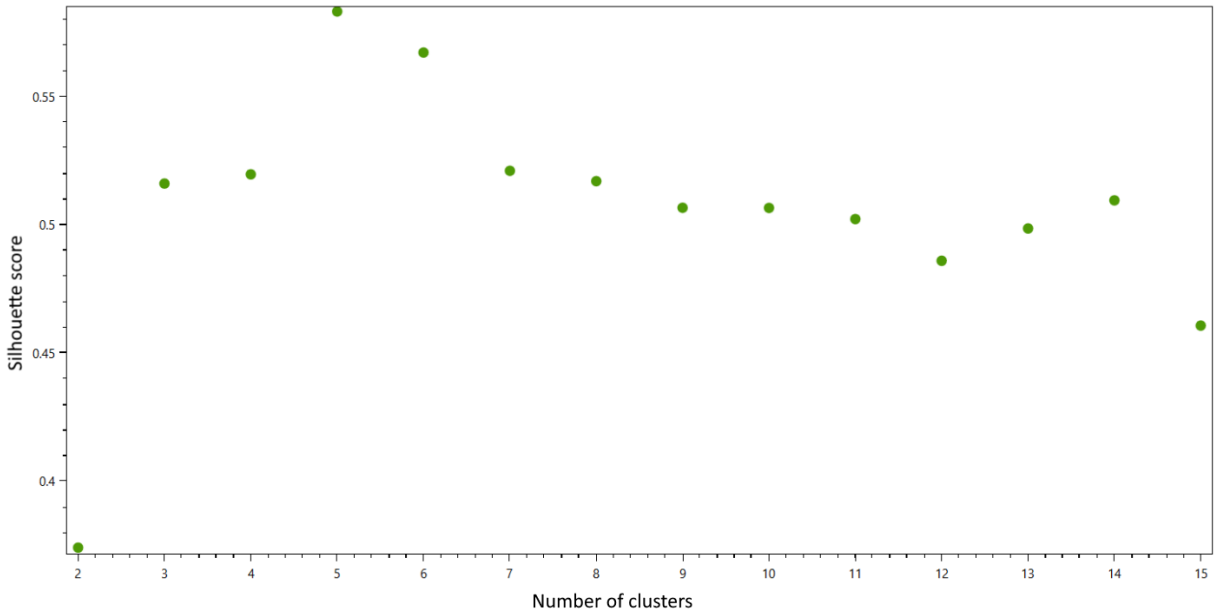


Figure 5-5 – Silhouette scores for various clustering runs. Result of the tests made by the clustering algorithm with different number of clusters (*k*) and their respective Silhouette scores. For the HSP90 experiment, in this particular run of the algorithm, the highest score was for *k* = 5.

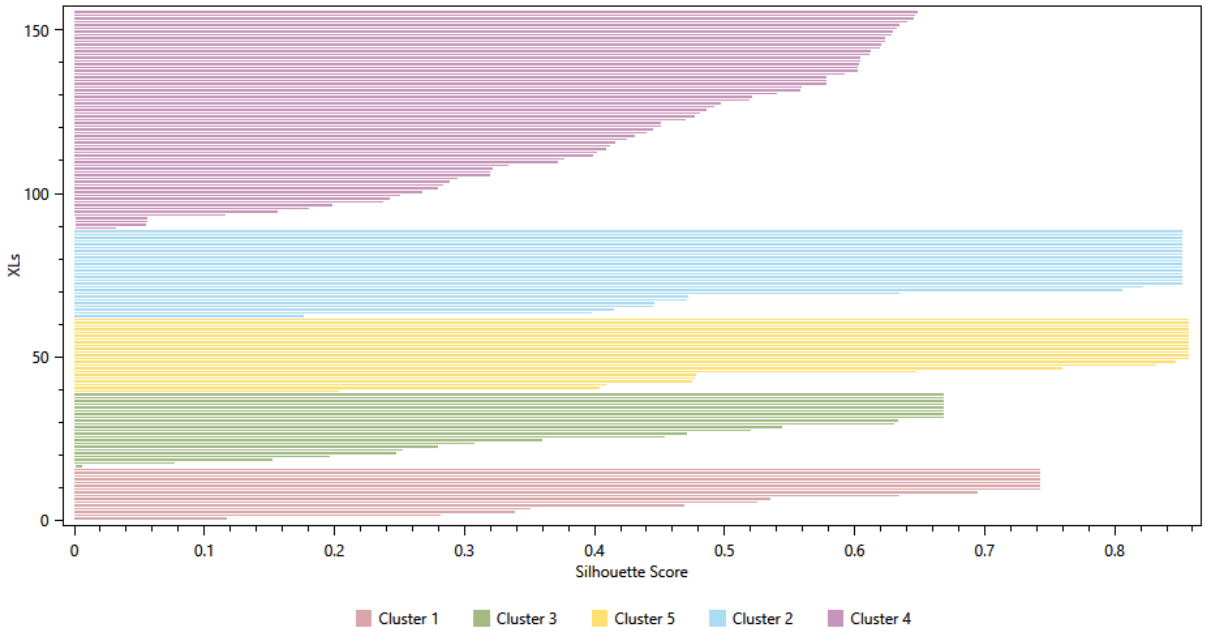
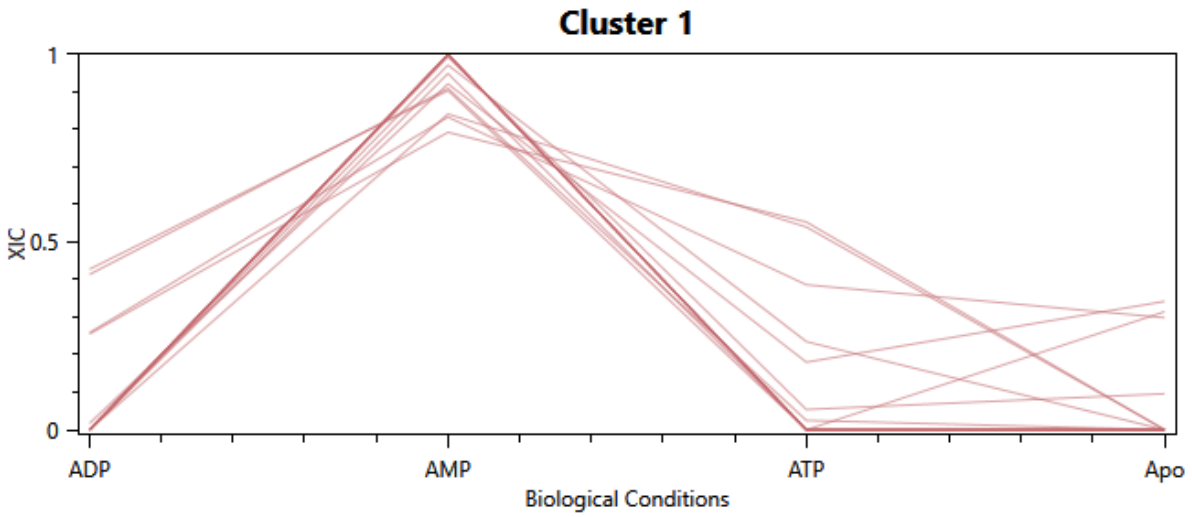
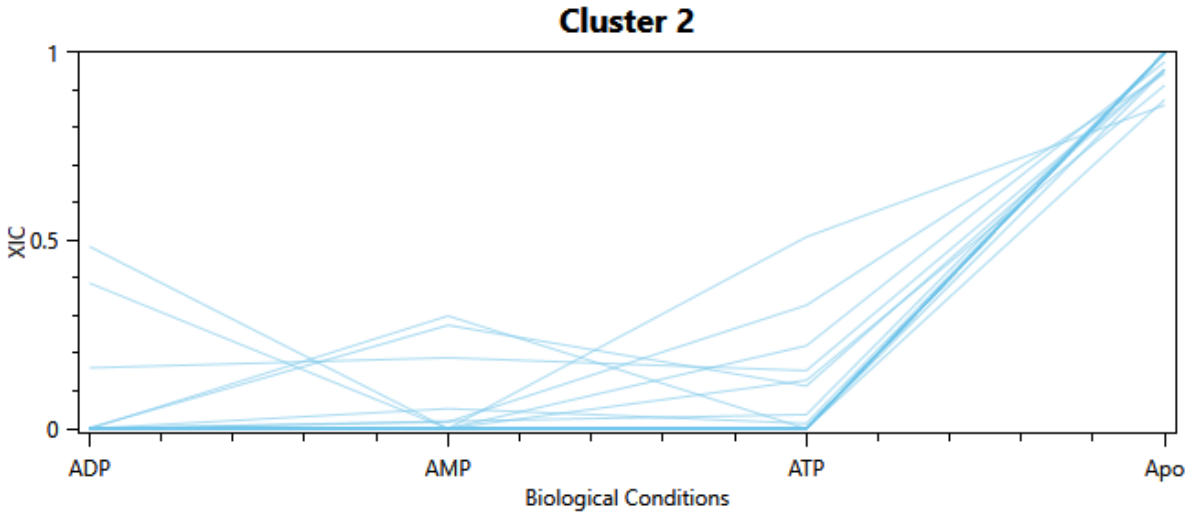
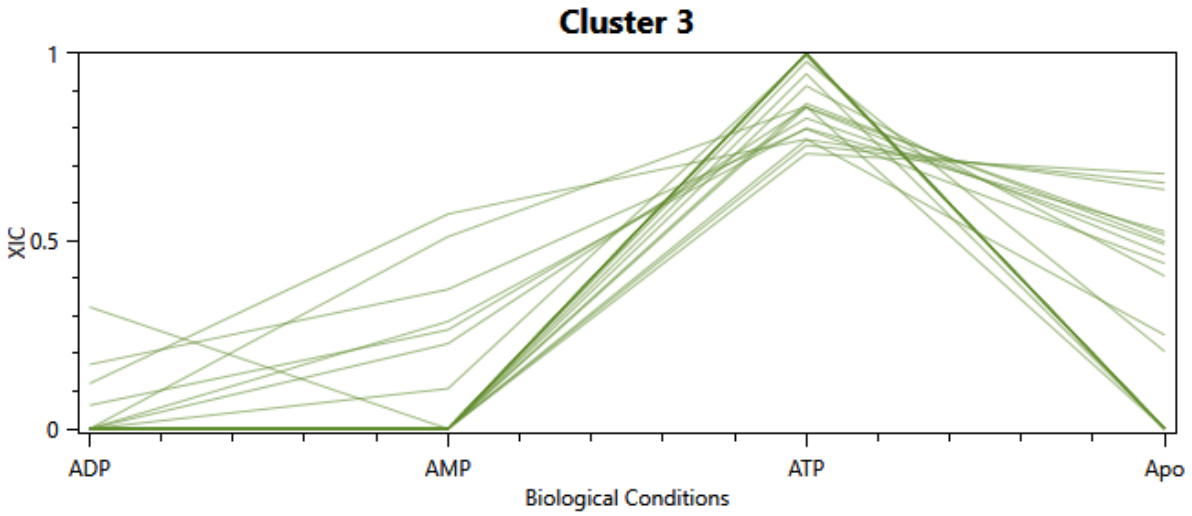


Figure 5-6 – Resulting Silhouette plot. Silhouette plot generated for the five clusters found by QUIN-XL. Here, each horizontal bar represents a XL within a cluster (which is colour coded), with its respective score. XLs with scores closer to 1 are in the right cluster; closer to 0 means they are in a grey zone between clusters and negative scores indicate that the XL probably does not belong to that cluster. No XL had a negative score in this run, indicating that all are assigned to the correct cluster, or in a grey zone.

The five clusters generated were the following (**Figure 5-7**):


Cluster 1
1
0.5
0
XIC
ADP
AMP
ATP
Apo
Biological Conditions


Cluster 2
1
0.5
0
XIC
ADP
AMP
ATP
Apo
Biological Conditions


Cluster 3
1
0.5
0
XIC
ADP
AMP
ATP
Apo
Biological Conditions

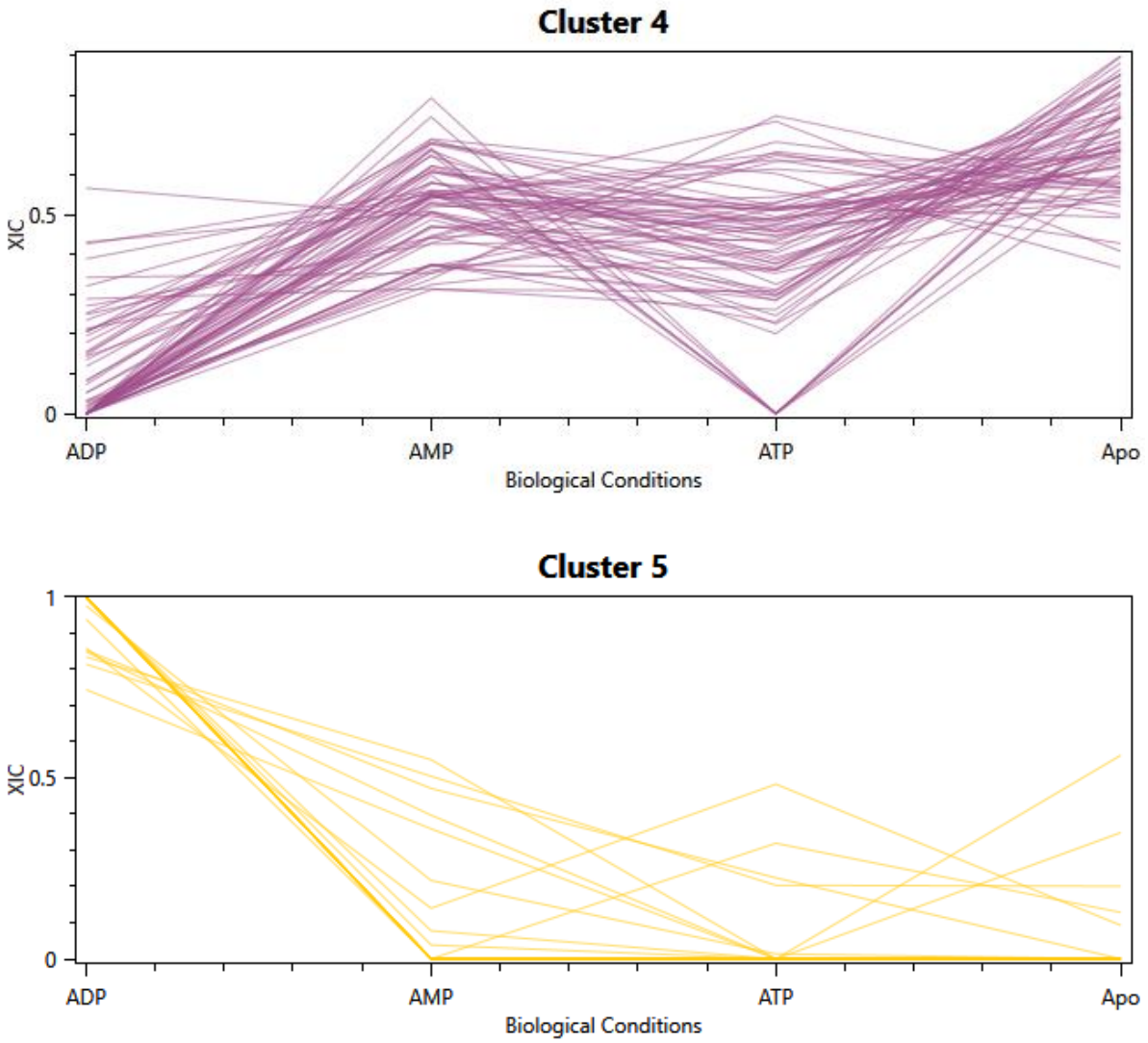


Figure 5-7 - The five clusters generated by QUIN-XL. Cluster 1 contains mostly XLs predominant in the AMP condition, Cluster 2 in the Apo condition, Cluster 3 in the ATP condition, with Cluster 4 showing XLs shared by all other conditions with a similar quantitative pattern. Cluster 5 has most of its XLs in the ADP condition.

We can see that four of the clusters (1, 2, 3, and 5) had cross-links mostly exclusive or more abundant in one condition, while Cluster 4 showed XLs shared by the four conditions and following a similar pattern. Some attempts of reclustering groups 1, 2, 3, and 5 usually ended up with a worse score or with clusters with only one or two XLs, so to avoid cases of overfitting they were analysed as they are shown here. Cluster 4 however, which was heavily populated by a variety of different quantitative-profiles, proved more informative when split into two new clusters, which are shown in **Figure 5-8**. One of the new clusters, Cluster 6, includes XLs with a similar pattern of being of low abundance in the ADP condition and increasing in the others, with the highest abundance being in the Apo state.

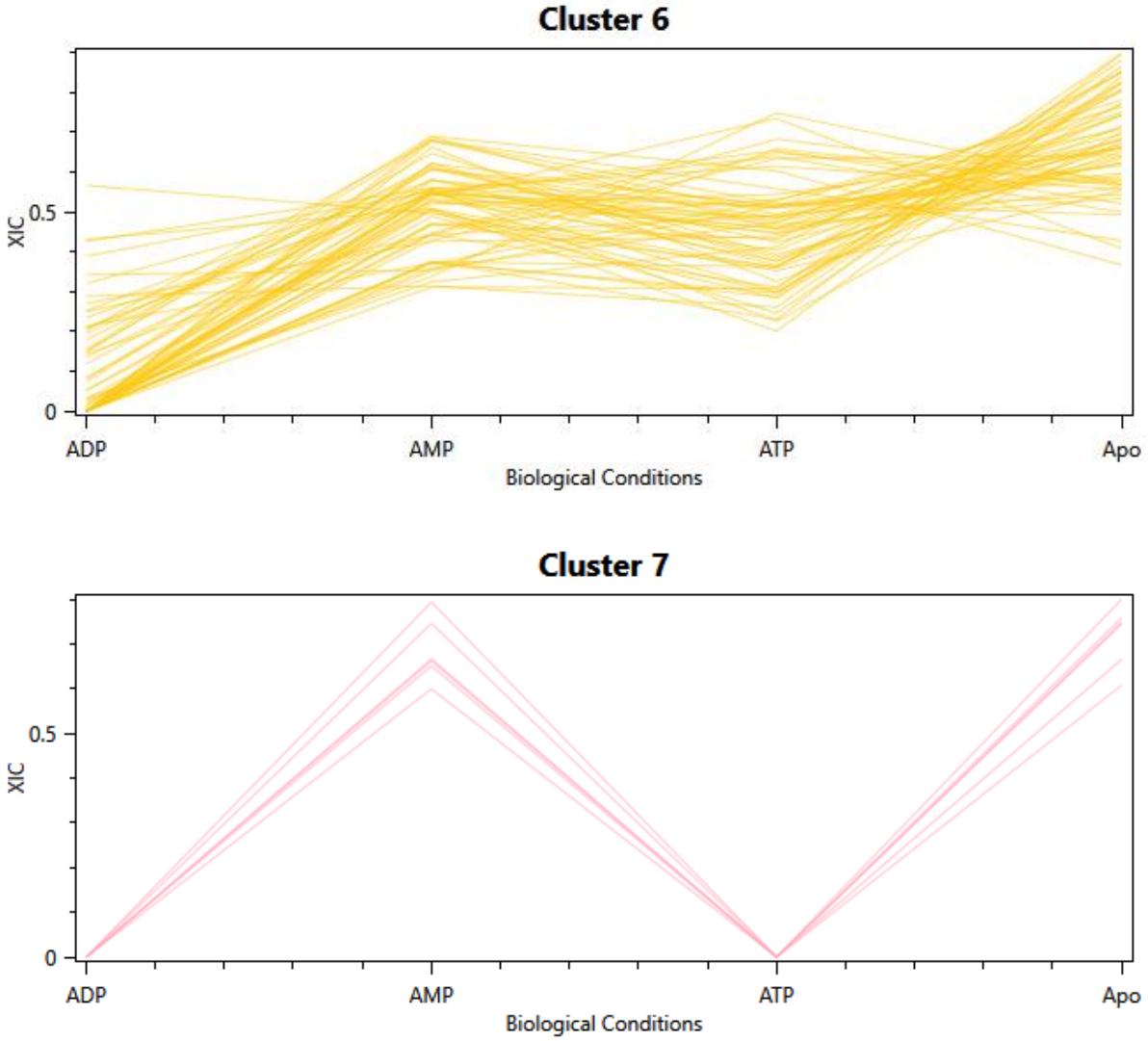


Figure 5-8 - Groups generated from reclustering former Cluster 4 into Clusters 6 and 7. We can see now that Cluster 6 shows a more similar pattern between the XLs, while Cluster 7 shows XLs coming from conditions AMP and Apo.

This way, with the XLs successfully quantified and clustered, it is easier to look at specific cross-links for further analysis, without having to manually separate them. The XLs in each cluster can be informative on the conformer they represent, such as the regions that differ the most amongst them by looking at the enriched XLs in each condition. Since there are four clusters with XLs mostly exclusive to a condition, we can assume that there are at least four different conformers in the samples, with the two remaining clusters representing structurally conserved regions in the protein. It is worth noting again that HSP90 is being used to validate the software, so we can assume what conformer is supposed to be most abundant in each condition, but that might not always be the case in other experiments.

In the next sections we will analyse each cluster separately to address structural variations, while also confronting it with crystallographic information. For that, models of the whole structure of HSP90 were created for the conditions Apo (no ligand, open structure), ATP-bound (closed conformation) and ADP-bound (closed conformation). No model for the AMP-bound condition was made as there are no crystallographic structures to serve as source.

The models were created using the I-TASSER server [67], which gets the PDB structure with the best sequence alignment and builds the new structure using that as the basis. This approach was chosen because, while there are crystallographic structures for HSP90 in databases, they are always of the separate domains, so in order to make the analysis easier, a model of the whole structure was preferred.

5.3.1 Apo state

The HSP90 Apo conformer presents an open structure and no ligand bound to it. We expect it to be represented by Cluster 2, as this is where the XLs most abundant to that condition appeared. With the exception of Cluster 6, which is derived from several shared XLs amongst the conditions, Cluster 2 is the most populated one, which was expected as, having a more open structure, the amino acids in the Apo conformer are more available to the solvent, facilitating the creation of cross-links.

By looking at the 2D-Map of Cluster 2 (**Figure 5-9**), it is possible to see that most XLs come from the middle-domain, with a few starting at N-terminal domain, which is expected as, since there is no ligand, the nucleotide binding pocket is likely to be more exposed. There are very few cross-links from the C-terminal domain, which might be explained by it being the site of dimerization, leaving little room for links to be formed.

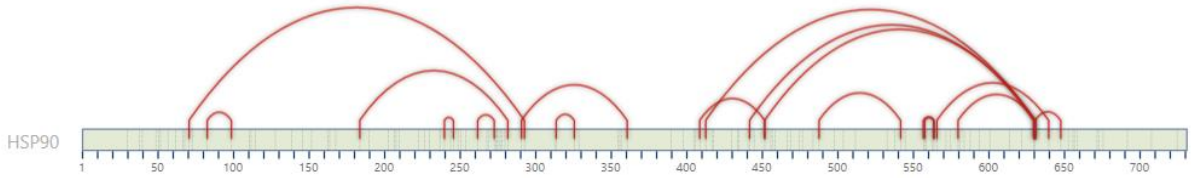


Figure 5-9 – 2D-Map from Cluster 2. 2D-Map of the links (represented by red arches) found in Cluster 2 for the HSP90 protein. The middle domain goes from position 272 to 629.

The 2D-Map of Cluster 7 (**Figure 5-10**), which contains links shared by the conditions Apo and AMP, shows a similar distribution of cross-links, again indicating that the N-terminal domain is probably exposed in these conformations.

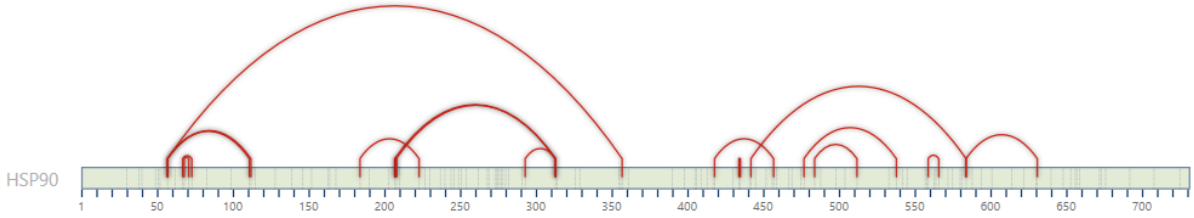


Figure 5-10 - 2D-Map with the cross-links identified in Cluster 7.

Comparing the information given by the clusters with the 3D model of the Apo conformation (**Figure 5-11**), it is possible to see that some XLs appear between amino acids structurally far apart, suggesting that the Apo state is very fluid, having more than one possible conformation, which was also shown in a paper by Chavez at al [68], where they found a more compact Apo conformation, with the N-terminal domain folding towards the middle domain. These cross-links could be used to generate new models for the Apo state of the HSP90 protein.

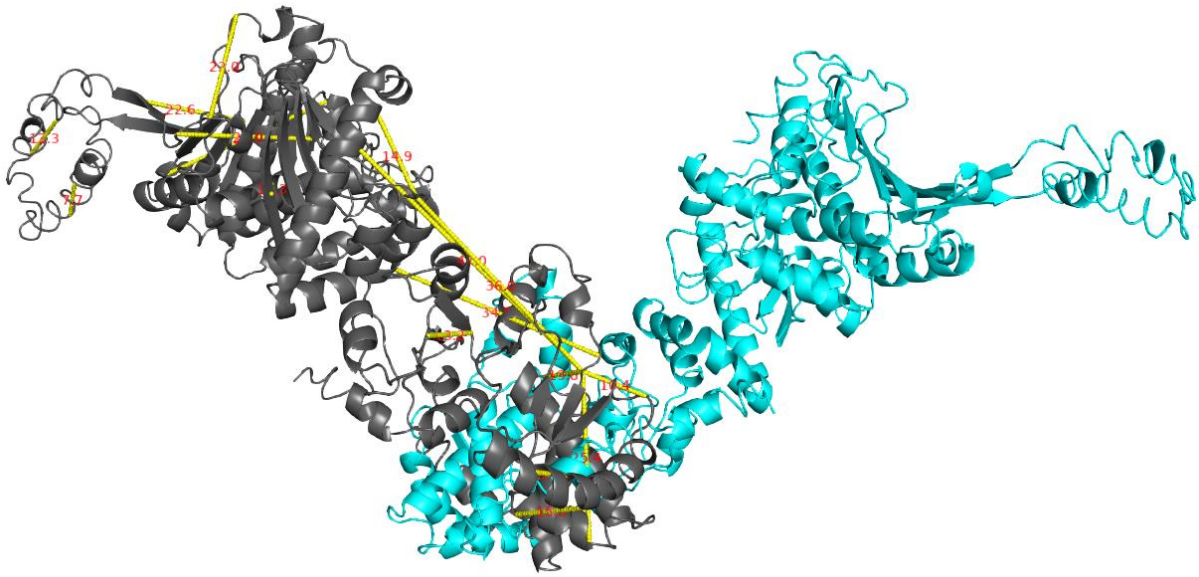


Figure 5-11 - 3D representation of the Apo state model. In grey and cyan, the two monomers of the HSP90, and in yellow the cross-links enriched in this conformation.

5.3.2 AMP-bound condition

The AMP-bound state does not have a crystallographic structure described in the literature; therefore, we cannot validate the cluster with that approach. It is possible however, to derive several hints regarding its structure based on the clusters generated by QUIN-XL.

Cluster 1 is the one where the XLs are exclusive or more abundant in the AMP condition, so it is the one we will analyse here. The clusters in the 2D-Map (**Figure 5-12**) indicate that most of the cross-links come from the N-terminal domain, suggesting that the AMP-bound conformer has a more exposed nucleotide binding pocket.

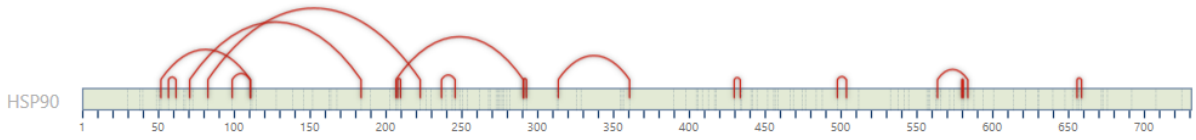


Figure 5-12 - 2D-Map representation of the XLs grouped in Cluster 1.

Given that Cluster 7 shares XLs with the AMP and Apo conditions, and these two states have mostly the same quantitative pattern in Cluster 6, it is reasonable to hypothesize that these two conditions have similar conformers, which is indicative that the nucleotide AMP does not bind very effectively or for very long to the HSP90 protein, shifting it to a more open conformation.

5.3.3 ADP-bound condition

The HSP90 ADP-bound conformer shows a closed conformation after the hydrolysis of the nucleotide ATP and, because of that, will probably have fewer amino acids available to the solvent, resulting in a lower number of XLs identified. That is exactly what happened with Cluster 5, which is the one with XLs exclusive or most abundant in the ADP-bound condition.

A look at the 2D-Map of XLs from Cluster 5 (**Figure 5-13**) shows that this condition has mostly intralinks in the middle to C-terminal domains, confirming the difficulty for the cross-linker to reach all the sites, and showing that some changes occur in that region in order to approximate the two monomers.

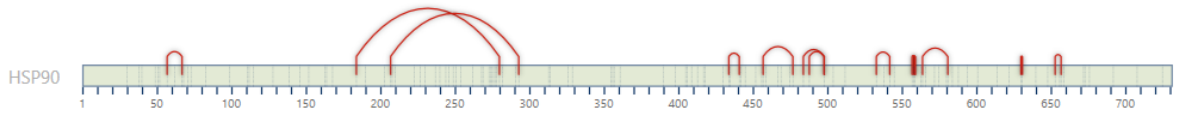


Figure 5-13 – 2D-Map from Cluster 5. The 2D-Map of XLs from Cluster 5, showing mostly intralinks in the middle-domain of HSP90.

Looking at the 3D model for the ADP model (**Figure 5-14**), it is possible to see that there are indeed some changes in the C-terminal domain when compared to the Apo (open) conformation (**Figure 5-11**).

Figure 5-14 - 3D model for the ADP closed conformation. In green and cyan are the two monomers of the HSP90, while the structure in magenta highlights the C-terminal domain of the green chain; in yellow are XLs identified.

5.3.4 ATP-bound condition

The ATP-bound HSP90 conformer presents a closed conformation, due to the dimerization of the N-terminal domain after the nucleotide is bound. In that case, very few links in the N-terminal domain would be expected, as it should not be very exposed. This is confirmed when looking at the 2D-Map of Cluster 3 (**Figure 5-15**), which shows XLs exclusive or most abundant in condition ATP.

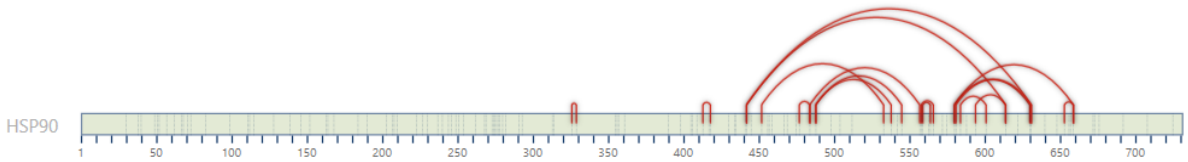


Figure 5-15 – 2D-Map from Cluster 3. 2D-Map of XLs from Cluster 3, which had mostly exclusive XLs from the ATP condition.

The 2D-Map for Cluster 3 shows that no XL exclusive from the ATP condition was formed in the N-terminal region, as anticipated. It also shows that most differences come from the C-terminal domain, which, just as in the ADP conformation, suffers some changes in order to close the monomers. The 3D model in **Figure 5-16** compared with the other two models corroborates this.

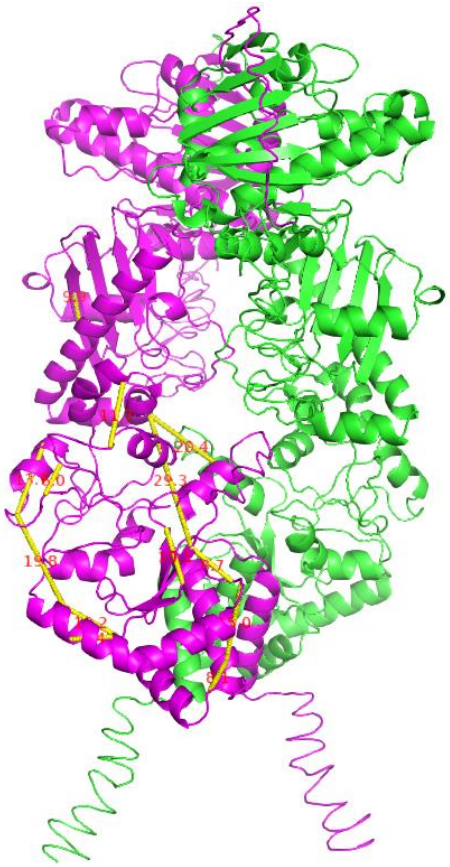


Figure 5-16 - 3D model of the ATP closed structure. In magenta and green are the two monomers, while in yellow are the XLs enriched in this conformation.

6. CONCLUSION AND PERSPECTIVES

Here we presented RawVegetable and QUIN-XL, our contributions toward quality control of mass spectrometry experiments and quantitative clustering of cross-linking mass spectrometry data, respectively.

We demonstrated the effectiveness of RawVegetable by assessing the quality of an XL-MS experiment. We developed a strategy for contiguous analysis of chromatography patterns with the distribution of spectra in the run, along with other informative features for interpreting the data. This made possible to plan for new experiments and identified problematic runs. We highlight the less populated HSP90 ATP_1 replicate (Section **5.2**), and the XL-MS gradient optimizations suggested by the far higher density of XL identifications at chromatography retention times 20 to 35 minutes. Both insights were made possible using RawVegetable.

QUIN-XL was able to apply label-free quantification strategies, such as XIC, both for sake of quantifying XL events, as well as for obtaining quantitative profiles for XL throughout different experiments. The software is also able to cluster the XLs identified and list groups of enriched species at different biological conditions, as per the experimental plan. These clusters, with the aid of QUIN-XL's graphical interpretation, can be used to provide information of which regions of the protein show more modifications in different biological states and conformations. QUIN-XL also has full exports for working alongside important tools in the field of structural protein analysis, such as PyMOL and Topolink.

With this methodology we were able to draw conclusions about the HSP90 experiment and have some understanding of the various conformations of the protein. One of the most interesting results was regarding the conformation for the HSP90 bound to the nucleotide AMP, which does not have a resolved crystallographic structure. QUIN-XL enabled hypothesizing that this state may have an open conformation very similar to that of the Apo state, which probably means that AMP does not bind well to the HSP90. Another interesting insight was about the flexibility of the Apo state, confirming information from the literature that other conformations exist for that condition aside from the ones already in crystallographic form.

To our knowledge, QUIN-XL is the first tool to apply a quantitative XL-MS methodology specifically to characterize protein conformers. We plan to keep working with XL-MS data with the objective of eventually creating a single environment for the analysis of cross-linked peptides, from identification, quantification, interaction

mapping, and structural analysis. QUIN-XL and RawVegetable will be important parts of this environment along with other solutions developed by our group.

REFERENCES

[1] N.H.C.S. Silva, C. Vilela, I.M. Marrucho, C.S.R. Freire, C. Pascoal Neto, A.J.D. Silvestre, Protein-based materials: from sources to innovative sustainable materials for biomedical applications, J. Mater. Chem. B. 2 (2014) 3715. https://doi.org/10.1039/c4tb00168k.

[3] Breda, A, Valadares, NF, Norberto de Souza, O, Garratt, RC, Chapter A06 - Protein Structure, Modelling and Applications, in: Bioinforma. Trop. Dis. Res. Pract. Case-Study Approach, Arthur Gruber, Alan M Durham, Chuong Huynh, and Hernando A del Portillo., Bethesda (MD): National Center for Biotechnology Information (US), 2008. https://www.ncbi.nlm.nih.gov/books/NBK6824/ (accessed December 18, 2019).

[4] C.V. Robinson, A. Sali, W. Baumeister, The molecular sociology of the cell, Nature. 450 (2007) 973–982. https://doi.org/10.1038/nature06523.

[5] H. Bai, W. Ma, S. Liu, L. Lai, Dynamic property is a key determinant for protein–protein interactions, Proteins Struct. Funct. Bioinforma. 70 (2008) 1323–1331. https://doi.org/10.1002/prot.21625.

[6] P. Kursula, The many structural faces of calmodulin: a multitasking molecular jackknife, Amino Acids. 46 (2014) 2295–2304. https://doi.org/10.1007/s00726-014-1795-y.

[7] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515. https://doi.org/10.1093/nar/gky1049.

[8] H.M. Berman, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242. https://doi.org/10.1093/nar/28.1.235.

[9] H. Deng, Y. Jia, Y. Zhang, Protein structure prediction, Int. J. Mod. Phys. B. 32 (2018) 1840009. https://doi.org/10.1142/S021797921840009X.

[10] J.D. Chavez, N.L. Liu, J.E. Bruce, Quantification of Protein−Protein Interactions with Chemical Cross-Linking and Mass Spectrometry, J. Proteome Res. 10 (2011) 1528–1537. https://doi.org/10.1021/pr100898e.

[11] A. Sinz, C. Arlt, D. Chorev, M. Sharon, Chemical cross-linking and native mass spectrometry: A fruitful combination for structural biology, Protein Sci. 24 (2015) 1193–1209. https://doi.org/10.1002/pro.2696.

[12] A. Sinz, Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions, Mass Spectrom. Rev. 25 (2006) 663–682. https://doi.org/10.1002/mas.20082.

[13] M. Fioramonte, H.C.R. de Jesus, A.J.R. Ferrari, D.B. Lima, R.L. Drekener, C.R.D. Correia, L.G. Oliveira, A.G. da C. Neves-Ferreira, P.C. Carvalho, F.C. Gozzo, XPlex: An Effective, Multiplex Cross-Linking Chemistry for Acidic Residues, Anal. Chem. 90 (2018) 6043–6050. https://doi.org/10.1021/acs.analchem.7b05135.

[14] J.T. Melchior, R.G. Walker, J. Morris, M.K. Jones, J.P. Segrest, D.B. Lima, P.C. Carvalho, F.C. Gozzo, M. Castleberry, T.B. Thompson, W.S. Davidson, An Evaluation of the Crystal Structure of C-terminal Truncated Apolipoprotein A-I in Solution Reveals Structural Dynamics Related to Lipid Binding, J. Biol. Chem. 291 (2016) 5439–5451. https://doi.org/10.1074/jbc.M115.706093.

[15] Z. Kukacka, M. Rosulek, M. Strohalm, D. Kavan, P. Novak, Mapping protein structural changes by quantitative cross-linking, Methods. 89 (2015) 112–120. https://doi.org/10.1016/j.ymeth.2015.05.027.

[16] F.J. O'Reilly, J. Rappsilber, Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology, Nat. Struct. Mol. Biol. 25 (2018) 1000–1008. https://doi.org/10.1038/s41594-018-0147-0.

[17] A. Leitner, M. Faini, F. Stengel, R. Aebersold, Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines, Trends Biochem. Sci. 41 (2016) 20–32. https://doi.org/10.1016/j.tibs.2015.10.008.

[18] A.N. Calabrese, S.E. Radford, Mass spectrometry-enabled structural biology of membrane proteins, Methods. 147 (2018) 187–205. https://doi.org/10.1016/j.ymeth.2018.02.020.

[19] Thermo Scientific, Crosslinking Technical Handbook, (2012). https://tools.thermofisher.com/content/sfs/brochures/1602163-Crosslinking-Reagents-Handbook.pdf (accessed December 31, 2019).

[20] A. Leitner, T. Walzthoeni, A. Kahraman, F. Herzog, O. Rinner, M. Beck, R. Aebersold, Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics, Mol. Cell. Proteomics. 9 (2010) 1634–1649. https://doi.org/10.1074/mcp.R000001-MCP201.

[21] A. Leitner, T. Walzthoeni, R. Aebersold, Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline, Nat. Protoc. 9 (2014) 120–137. https://doi.org/10.1038/nprot.2013.168.

[22] S. Mädler, C. Bich, D. Touboul, R. Zenobi, Chemical cross-linking with NHS esters: a systematic study on amino acid reactivities, J. Mass Spectrom. 44 (2009) 694–706. https://doi.org/10.1002/jms.1544.

[23] C.L. Swaim, J.B. Smith, D.L. Smith, Unexpected products from the reaction of the synthetic cross-linker 3,3′-dithiobis(sulfosuccinimidyl propionate), DTSSP with peptides, J. Am. Soc. Mass Spectrom. 15 (2004) 736–749. https://doi.org/10.1016/j.jasms.2004.01.011.

[24] E.D. Merkley, J.R. Cort, J.N. Adkins, Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze, J. Struct. Funct. Genomics. 14 (2013) 77–90. https://doi.org/10.1007/s10969-013-9160-z.

[25] A. Sinz, The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks, Expert Rev. Proteomics. 11 (2014) 733–743. https://doi.org/10.1586/14789450.2014.960852.

[26] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature. 422 (2003) 198–207. https://doi.org/10.1038/nature01511.

[27] J.R. Yates, C.I. Ruse, A. Nakorchevsky, Proteomics by Mass Spectrometry: Approaches, Advances, and Applications, Annu. Rev. Biomed. Eng. 11 (2009) 49–79. https://doi.org/10.1146/annurev-bioeng-061008-124934.

[28] J.J. Pitt, Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry, Clin. Biochem. Rev. 30 (2009) 19–34.

[29] B. Canas, Mass spectrometry technologies for proteomics, Brief. Funct. Genomic. Proteomic. 4 (2006) 295–320. https://doi.org/10.1093/bfgp/eli002.

[30] Michael. Karas, Franz. Hillenkamp, Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons, Anal. Chem. 60 (1988) 2299–2301. https://doi.org/10.1021/ac00171a028.

[31] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, C.M. Whitehouse, Electrospray ionization for mass spectrometry of large biomolecules, Science. 246 (1989) 64–71. https://doi.org/10.1126/science.2675315.

[32] L. Konermann, E. Ahadi, A.D. Rodriguez, S. Vahidi, Unraveling the Mechanism of Electrospray Ionization, Anal. Chem. 85 (2013) 2–9. https://doi.org/10.1021/ac302789c.

[33] C.S. Ho, C.W.K. Lam, M.H.M. Chan, R.C.K. Cheung, L.K. Law, L.C.W. Lit, K.F. Ng, M.W.M. Suen, H.L. Tai, Electrospray ionisation mass spectrometry: principles and clinical applications, Clin. Biochem. Rev. 24 (2003) 3–12.

[34] J.P. Savaryn, T.K. Toby, N.L. Kelleher, A researcher's guide to mass spectrometry-based proteomics, PROTEOMICS. 16 (2016) 2435–2443. https://doi.org/10.1002/pmic.201600113.

[35] B. Domon, Mass Spectrometry and Protein Analysis, Science. 312 (2006) 212–217. https://doi.org/10.1126/science.1124619.

[36] J.K. Eng, A.L. McCormack, J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom. 5 (1994) 976–989. https://doi.org/10.1016/1044-0305(94)80016-2.

[37] D. Borges, Y. Perez-Riverol, F.C.S. Nogueira, G.B. Domont, J. Noda, F. da Veiga Leprevost, V. Besada, F.M.G. França, V.C. Barbosa, A. Sánchez, P.C. Carvalho, Effectively addressing complex proteomic search spaces with peptide spectrum matching, Bioinforma. Oxf. Engl. 29 (2013) 1343–1344. https://doi.org/10.1093/bioinformatics/btt106.

[38] R.E. Higgs, J.P. Butler, B. Han, M.D. Knierman, Quantitative Proteomics via High Resolution MS Quantification: Capabilities and Limitations, Int. J. Proteomics. 2013 (2013) 1–10. https://doi.org/10.1155/2013/674282.

[39] K. Aoshima, K. Takahashi, M. Ikawa, T. Kimura, M. Fukuda, S. Tanaka, H.E. Parry, Y. Fujita, A.C. Yoshizawa, S. Utsunomiya, S. Kajihara, K. Tanaka, Y. Oda, A simple peak detection and label-free quantitation algorithm for chromatography-mass spectrometry, BMC Bioinformatics. 15 (2014). https://doi.org/10.1186/s12859-014-0376-0.

[40] S.E. Jackson, Hsp90: Structure and Function, in: S. Jackson (Ed.), Mol. Chaperones, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 155–240. https://doi.org/10.1007/128_2012_356.

[41] P. Csermely, T. Schnaider, C. So″ti, Z. Prohászka, G. Nardai, The 90-kDa Molecular Chaperone Family, Pharmacol. Ther. 79 (1998) 129–168. https://doi.org/10.1016/S0163-7258(98)00013-8.

[42] A. Hoter, M.E. El-Sabban, H.Y. Naim, The HSP90 Family: Structure, Regulation, Function, and Implications in Health and Disease, Int. J. Mol. Sci. 19 (2018). https://doi.org/10.3390/ijms19092560.

[43] J.C. Young, I. Moarefi, F.U. Hartl, Hsp90, J. Cell Biol. 154 (2001) 267–274. https://doi.org/10.1083/jcb.200104079.

[44] K.A. Krukenberg, T.O. Street, L.A. Lavery, D.A. Agard, Conformational dynamics of the molecular chaperone Hsp90, Q. Rev. Biophys. 44 (2011) 229–255. https://doi.org/10.1017/S0033583510000314.

[45] M. Amaral, D.B. Kokh, J. Bomke, A. Wegener, H.P. Buchstaller, H.M. Eggenweiler, P. Matias, C. Sirrenberg, R.C. Wade, M. Frech, Protein conformational flexibility modulates kinetics and thermodynamics of drug binding, Nat. Commun. 8 (2017) 2276. https://doi.org/10.1038/s41467-017-02258-w.

[46] P. Wortmann, M. Götz, T. Hugel, Cooperative Nucleotide Binding in Hsp90 and Its Regulation by Aha1, Biophys. J. 113 (2017) 1711–1718. https://doi.org/10.1016/j.bpj.2017.08.032.

[47] H.M. Kumalo, S. Bhakat, M.E. Soliman, Heat-Shock Protein 90 (Hsp90) as Anticancer Target for Drug Discovery: An Ample Computational Perspective, Chem. Biol. Drug Des. 86 (2015) 1131–1160. https://doi.org/10.1111/cbdd.12582.

[48] S.Y. Hyun, H.T. Le, C.-T. Nguyen, Y.-S. Yong, H.-J. Boo, H.J. Lee, J.-S. Lee, H.-Y. Min, J. Ann, J. Chen, H.-J. Park, J. Lee, H.-Y. Lee, Development of a novel Hsp90 inhibitor NCT-50 as a potential anticancer

agent for the treatment of non-small cell lung cancer, Sci. Rep. 8 (2018) 13924. https://doi.org/10.1038/s41598-018-32196-6.

[49] D.B. Lima, T.B. de Lima, T.S. Balbuena, A.G.C. Neves-Ferreira, V.C. Barbosa, F.C. Gozzo, P.C. Carvalho, SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis, J. Proteomics. (2015). https://doi.org/10.1016/j.jprot.2015.01.013.

[50] P.C. Carvalho, D.B. Lima, F.V. Leprevost, M.D.M. Santos, J.S.G. Fischer, P.F. Aquino, J.J. Moresco, J.R. Yates, V.C. Barbosa, Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0, Nat. Protoc. 11 (2015) 102–117. https://doi.org/10.1038/nprot.2015.133.

[51] P.C. Carvalho, T. Xu, X. Han, D. Cociorva, V.C. Barbosa, J.R. Yates, YADA: a tool for taking the most out of high-resolution spectra, Bioinformatics. 25 (2009) 2734–2736. https://doi.org/10.1093/bioinformatics/btp489.

[52] D.B. Lima, J.T. Melchior, J. Morris, V.C. Barbosa, J. Chamot-Rooke, M. Fioramonte, T.A.C.B. Souza, J.S.G. Fischer, F.C. Gozzo, P.C. Carvalho, W.S. Davidson, Characterization of homodimer interfaces with cross-linking mass spectrometry and isotopically labeled proteins, Nat. Protoc. 13 (2018) 431–458. https://doi.org/10.1038/nprot.2017.113.

[53] B.O. Keller, J. Sui, A.B. Young, R.M. Whittal, Interferences and contaminants encountered in modern mass spectrometry, Anal. Chim. Acta. 627 (2008) 71–81. https://doi.org/10.1016/j.aca.2008.04.043.

[54] M. Rosenblatt, Remarks on Some Nonparametric Estimates of a Density Function, Ann. Math. Stat. 27 (1956) 832–837. https://doi.org/10.1214/aoms/1177728190.

[55] E. Parzen, On Estimation of a Probability Density Function and Mode, Ann. Math. Stat. 33 (1962) 1065–1076. https://doi.org/10.1214/aoms/1177704472.

[56] B.W. Silverman, Density estimation for statistics and data analysis, Chapman & Hall/CRC, Boca Raton, 1998.

[57] J.K. Patel, C.B. Read, Handbook of the normal distribution, 2nd ed., rev.expanded, Marcel Dekker, New York, 1996.

[58] S.H. Giese, A. Belsom, L. Sinn, L. Fischer, J. Rappsilber, Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses, Anal. Chem. 91 (2019) 2678–2685. https://doi.org/10.1021/acs.analchem.8b04037.

[59] T.O. Kvalseth, Cautionary Note about R 2, Am. Stat. 39 (1985) 279. https://doi.org/10.2307/2683704.

[60] H. Anton, Elementary linear algebra, 7th ed, John Wiley, New York, 1994.

[61] Abraham. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Anal. Chem. 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.

[62] H.H. Madden, Comments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data, Anal. Chem. 50 (1978) 1383–1386. https://doi.org/10.1021/ac50031a048.

[63] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory. 28 (1982) 129–137. https://doi.org/10.1109/TIT.1982.1056489.

[64] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

[65] A.J.R. Ferrari, M.A. Clasen, L. Kurt, P.C. Carvalho, F.C. Gozzo, L. Martínez, TopoLink: evaluation of structural models using chemical crosslinking distance constraints, Bioinformatics. 35 (2019) 3169–3170. https://doi.org/10.1093/bioinformatics/btz014.

[66] A. Kahraman, F. Herzog, A. Leitner, G. Rosenberger, R. Aebersold, L. Malmström, Cross-Link Guided Molecular Modeling with ROSETTA, PLoS ONE. 8 (2013) e73411. https://doi.org/10.1371/journal.pone.0073411.

[67] J. Yang, Y. Zhang, I-TASSER server: new development for protein structure and function predictions, Nucleic Acids Res. 43 (2015) W174–W181. https://doi.org/10.1093/nar/gkv342.

[68] J.D. Chavez, D.K. Schweppe, J.K. Eng, J.E. Bruce, In Vivo Conformational Dynamics of Hsp90 and Its Interactors, Cell Chem. Biol. 23 (2016) 716–726. https://doi.org/10.1016/j.chembiol.2016.05.012.

ANNEX 1

Journal of Proteomics 225 (2020) 103864

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

RawVegetable – A data assessment tool for proteomics and cross-linking mass spectrometry experiments

Louise U. Kurt[a,*], Milan A. Clasen[a], Marlon D.M. Santos[a], Tatiana A.C.B. Souza[a], Emanuella C. Andreassa[a], Eduardo B. Lyra[b], Diogo B. Lima[c], Fabio C. Gozzo[b], Paulo C. Carvalho[a,*]

[a] Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute, Fiocruz, Paraná, Brazil
[b] Institute of Chemistry, University of Campinas, São Paulo, Brazil
[c] Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

A B S T R A C T

We present RawVegetable, a software for mass spectrometry data assessment and quality control tailored toward shotgun proteomics and cross-linking experiments. RawVegetable provides four main modules with distinct features: (A) The charge state chromatogram that independently displays the ion current for each charge state; useful for optimizing the chromatography for highly charged ions and with lower XIC values such as those typically found in cross-linking experiments. (B) The XL-Artefact determination, which flags possible noncovalently associated peptides. (C) The TopN density estimation, for detecting retention time intervals of under or over-sampling, and (D) The chromatography reproducibility module, which provides pairwise comparisons between multiple experiments. RawVegetable, a tutorial, and the example data are freely available for academic use at: http://patternlabforproteomics.org/rawvegetable.

Significance: Chromatography optimization is a critical step for any shotgun proteomic or cross-linking mass spectrometry experiment. Here, we present a nifty solution with several key features, such as displaying individual charge state chromatograms, highlighting chromatographic regions of under- or over-sampling and checking for reproducibility.

1. Main text

Chromatography quality control (QC) is a critical step in any biological mass spectrometry experiment [1]. Several freely available tools tailored toward shotgun proteomics are available, such as: iMonDB [2], QCloud [3], rawDiag [4], and RawMeat (Vast Scientific); the latter being probably the most widely adopted. In analogy to RawMeat, we present RawVegetable, a nifty tool for general proteomics data assessment with a focus on cross-linking mass spectrometry (XLMS). RawVegetable includes all RawMeat QC features, handles other standard formats such as mzML, and presents several additional features presented below. We now highlight four of RawVegetable's features; a complete description of all the features is available in the project's website.

- **Charge state chromatogram**: XLMS emerged as a breakthrough for enabling large-scale protein-protein interaction studies [5] and structural proteomics [6]. In brief, XLMS comprises of the application of cross-linking reagents that covalently link to specific amino acids to ultimately provide distance constraints that aid in structural or protein-protein interaction experiments. Cross-linked peptides will mostly present higher charge states than the linear peptides found in traditional proteomic experiments, due to the existence of a second tryptic peptide in the molecule [7]. Moreover, cross-linked peptides are typically less abundant than linear ones, making optimization imperative. As such, chromatographic optimizations with existing tools are severely limited in the face of XLMS experiments as their viewers cannot independently account for highly charged peptide ions.

This charge state chromatogram module is tailored toward improving chromatography for highly charged ion species by independently plotting the chromatographic profile for each charge state (Fig. 1A). To achieve this, spectra are deconvoluted using the Y.A.D.A. algorithm [8]. RawVegetable can also read the SIM-XL [6] output file and plot where the cross-linked peptides have been identified.

- **XL-Artefact determination:** Recently, Giese et al. [9] described the presence of peptides associated noncovalently in XLMS experiments, leading to false inter-link identifications. RawVegetable flags these XL-Artefacts (Fig. 1A) from SIM-XL's results by looking for extracted ion chromatogram (XIC) curves of each peptide identified in the inter-link at the same retention time. A score is then calculated and assigned based on the identification of these XIC curves and their

* Corresponding authors.
E-mail addresses: lulrichkurt@gmail.com (L.U. Kurt), paulo@pcarvalho.com (P.C. Carvalho).

https://doi.org/10.1016/j.jprot.2020.103864
Received 9 July 2019; Received in revised form 29 April 2020; Accepted 3 June 2020
Available online 09 June 2020
1874-3919/

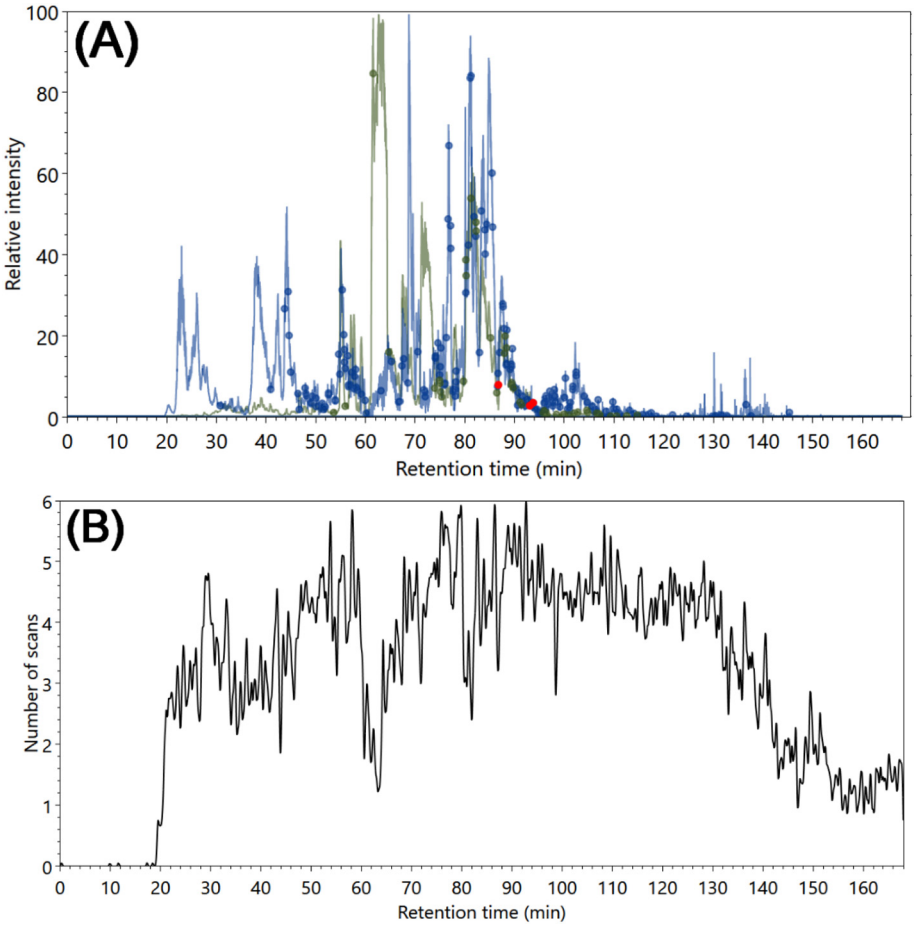


Fig. 1. (A) The charge 3+ (in green) and charge 4+ (in blue) chromatograms. The green and blue dots indicate where cross-linked species were identified with charges 3+ and 4+, respectively; the red dots indicate possible XL-artefacts that should be closely analyzed. (B) The number of MS/MS per duty cycle (TopN) versus time density estimation for the same chromatogram. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

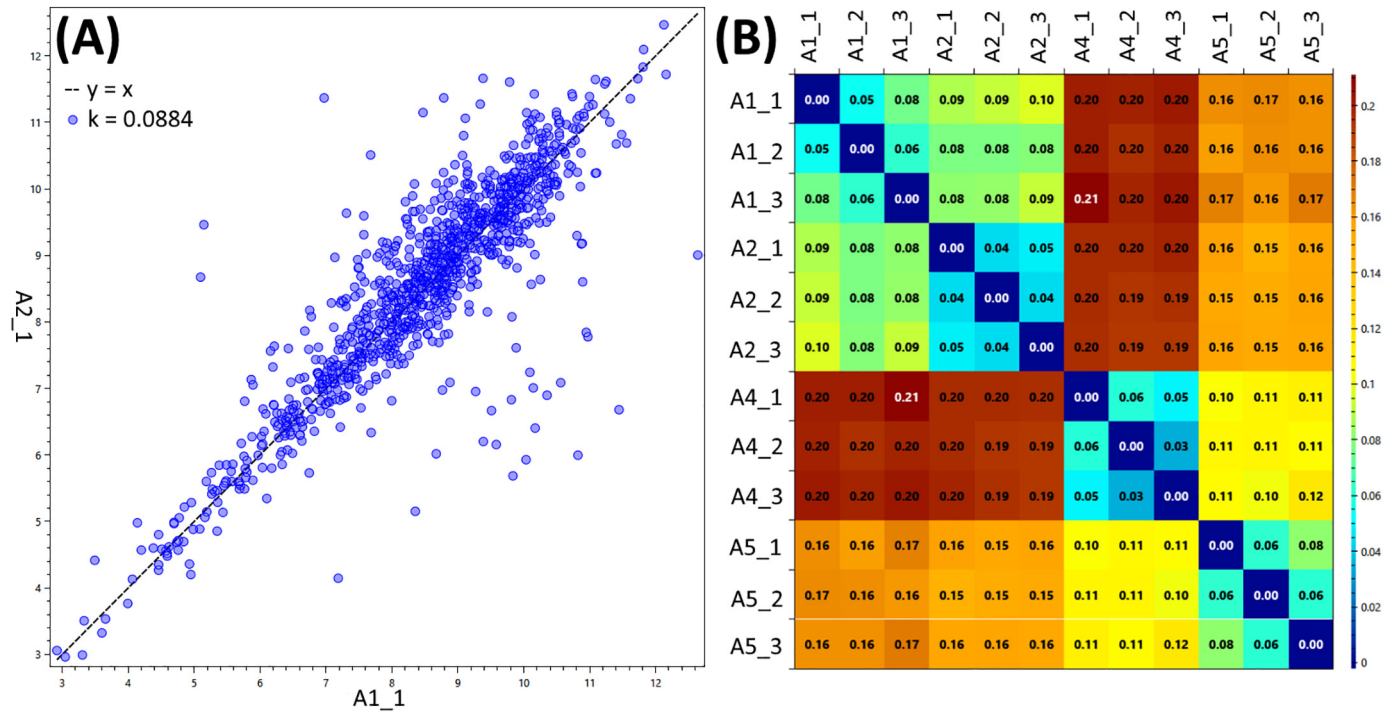


Fig. 2. (A) A comparison of XICs between two runs. The blue dots represent peptides found in both files; the axes are the logarithmic values for their XIC in each file. The *k* value is an adapted Euclidean distance metric, providing a measurement of how different the peptide XICs are for both runs. (B) The heatmap shows the *k* value for a comparison between all the files generated for the experiment. Regions with a blue colour and value closer to zero usually are comparisons between technical replicates, while regions between different conditions are expected to be red, as they differ the most. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

intensities.

- **TopN density estimation**: Our TopN distribution module was inspired by RawMeat's histogram of TopN spacing. However, this module provides an improved view for the density estimation of MS/MS scans per duty cycle along all the chromatographic retention time (Fig. 1B). This allows pin-pointing retention time intervals that could lead to over- or under-sampling, so the necessary gradient adjustments can be done. More information on how the density estimation is calculated can be found in the project website.

- **Chromatography reproducibility**: This module performs all pairwise comparisons of XICs of common peptides between identifications of different runs. To achieve this, PatternLab's [10] *.xic result, or the SIM-XL output files [6] serve as input to generate the dot plot (pairwise comparison) and heatmap (comparison among all runs) (Fig. 2). This module allows a bird's-eye view of the chromatographic reproducibility for all runs and facilitates quickly spotting problematic MS run.

Tables with general statistics such as number of MS and MS/MS spectra, a preliminary contaminant search, a histogram on fragmentation efficiency and a histogram of precursor's mass distribution are also available.

Data availability

http://proteomics.fiocruz.br/rawvegetable/data

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by Fiocruz, the Brazilian National Research Council (CNPq - Universal) and Graduate Studies Agency (CAPES).

References

[1] W. Bittremieux, D.L. Tabb, F. Impens, A. Staes, E. Timmerman, L. Martens, K. Laukens, Quality control in mass spectrometry-based proteomics, Mass Spectrom. Rev. 37 (2018) 697–711, https://doi.org/10.1002/mas.21544.
[2] W. Bittremieux, H. Willems, P. Kelchtermans, L. Martens, K. Laukens, D. Valkenborg, iMonDB: mass spectrometry quality control through instrument monitoring, J. Proteome Res. 14 (2015) 2360–2366, https://doi.org/10.1021/acs.jproteome.5b00127.
[3] C. Chiva, R. Olivella, E. Borràs, G. Espadas, O. Pastor, A. Solé, E. Sabidó, QCloud: a cloud-based quality control system for mass spectrometry-based proteomics laboratories, PLoS One 13 (2018) e0189209, , https://doi.org/10.1371/journal.pone.0189209.
[4] C. Trachsel, C. Panse, T. Kockmann, W.E. Wolski, J. Grossmann, R. Schlapbach, rawDiag - an R package supporting rational LC-MS method optimization for bottom-up proteomics, Bioinformatics (2018), https://doi.org/10.1101/304485.
[5] A. Sinz, Divide and conquer: cleavable cross-linkers to study protein conformation and protein-protein interactions, Anal. Bioanal. Chem. 409 (2017) 33–44, https://doi.org/10.1007/s00216-016-9941-x.
[6] D.B. Lima, J.T. Melchior, J. Morris, V.C. Barbosa, J. Chamot-Rooke, M. Fioramonte, T.A.C.B. Souza, J.S.G. Fischer, F.C. Gozzo, P.C. Carvalho, W.S. Davidson, Characterization of homodimer interfaces with cross-linking mass spectrometry and isotopically labeled proteins, Nat. Protoc. 13 (2018) 431–458, https://doi.org/10.1038/nprot.2017.113.
[7] Z.A. Chen, J. Rappsilber, Quantitative cross-linking/mass spectrometry to elucidate structural changes in proteins and their complexes, Nat. Protoc. 14 (2019) 171–201, https://doi.org/10.1038/s41596-018-0089-3.
[8] P.C. Carvalho, T. Xu, X. Han, D. Cociorva, V.C. Barbosa, J.R. Yates, YADA: a tool for taking the most out of high-resolution spectra, Bioinformatics. 25 (2009) 2734–2736, https://doi.org/10.1093/bioinformatics/btp489.
[9] S.H. Giese, A. Belsom, L. Sinn, L. Fischer, J. Rappsilber, Noncovalently associated peptides observed during liquid chromatography-mass spectrometry and their effect on cross-link analyses, Anal. Chem. 91 (2019) 2678–2685, https://doi.org/10.1021/acs.analchem.8b04037.
[10] P.C. Carvalho, D.B. Lima, F.V. Leprevost, M.D.M. Santos, J.S.G. Fischer, P.F. Aquino, J.J. Moresco, J.R. Yates, V.C. Barbosa, Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0, Nat. Protoc. 11 (2015) 102–117, https://doi.org/10.1038/nprot.2015.133.